Analyzing the Factors that Determine Housing Prices in Beijing in the 21st Century: Building Structure, Location, and Number of Living Rooms as Prime Drivers of Price Surges*

Xincheng Zhang

April 17, 2024

This report examines Beijing housing prices from 2011 to 2017, using Kaggle's beta API to collect data from Lianjia company. It focuses on average price per square meter as the main measure and analyzes factors like geographical location, building type, and building structure. The study identifies that building structure, location, and number of living rooms have the greatest impact on price increases. These findings could help predict housing price trends in Beijing and inform personal home-buying decisions and economic strategies.

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 reports 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 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. The estimand for this report is the average effect of certain factors such as geographical location coordinates, building type, number of kitchens, and other characteristics on housing prices in Beijing.

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. For the Model part, I will compare the impact of different housing factors on prices under distance distribution. The regression line equation and its impact on the interpretation will be described. Some image sketches are 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 the number of living rooms, location, 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 ID 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 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 ((0)/(1))

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: Features 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 are (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 created variables can be found in Table 3.

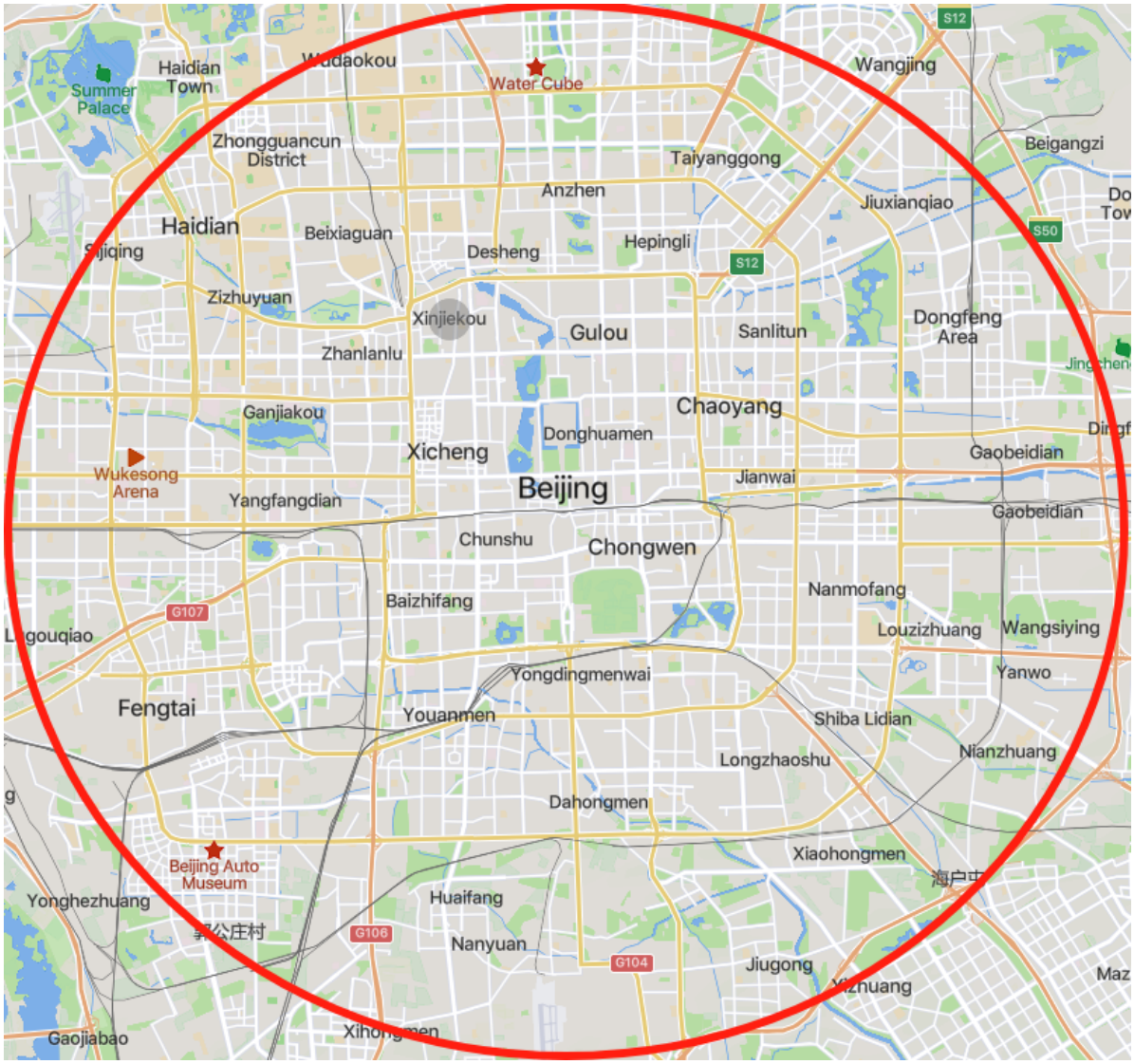


Figure 1: Beijing city center (39.901996392, 116.38833178) as the center of the circle

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 type of combination of plate and tower, which is also made of brick and wood in building structures, before 2000 and after 2000. These two dependencies will be explained in detail in the Result section.

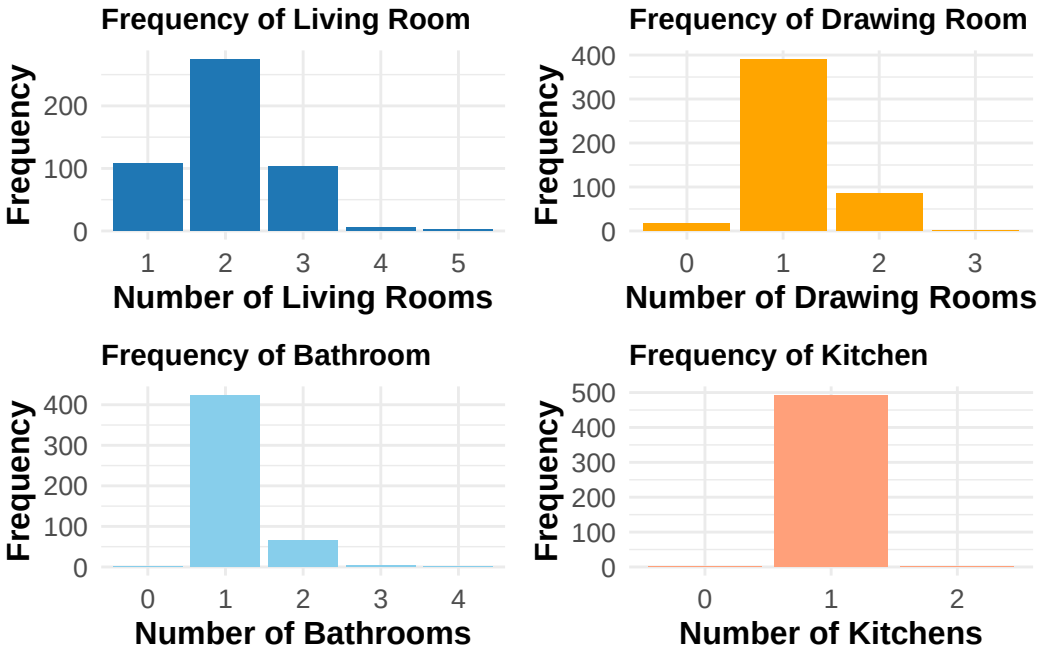


Figure 2: Summary statistics for different rooms

The measurements in this report mainly perform central tendency, dispersion, and summary statistics on the four-room type variables in the data set, including the number of living rooms, drawing rooms, bathrooms, and kitchens. Different measurements also inspired me to explore the impact of certain rooms on price increases. By Figure 2, for the cleaned data set with a total number of observations of 495, the number of living rooms appears the most in 2 listings, while the situations of 1 and 3 are the same, which makes me want to study more about the relationship between living room quantity and price. Similarly, listings with one drawing room also have a high degree of concentration compared to other situations, and the degree of dispersion is also appropriate. This distribution is also suitable for studying the relationship with price in subsequent work. On the contrary, summary statistics for bathroom and kitchen. Although it has the prominence of a single situation, it lacks dispersion. Such measurement performance makes statistical work meaningless since there are insufficient missing samples in each case. The information on residuals can be found in the Appendix.

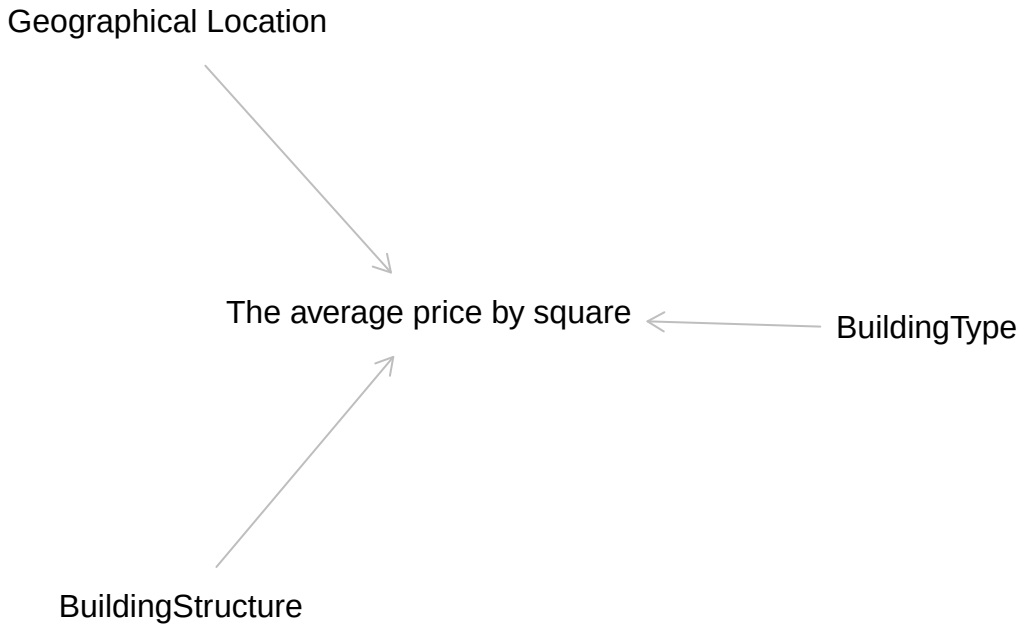
All data manipulated and presented in this report were sourced from the datasets (Ruiqurm 2024) in Kaggle open website. The data processing, and analyzing for this report is using R (R Core Team 2024a) along with other support packages tidyverse(Wickham et al. 2019), psych(William Revelle 2024), lubridate(Grolemund 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), gridExtra(Auguie 2017), arrow(Richardson et al. 2024), kableExtra(Zhu 2024).

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. 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 in the latitude and longitude coordinate system. 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 in order to enhance the precision of equations. 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, in RMB. β_0, β_0', β_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', β_1'' are the coefficients for the 'dif_cor' variable for each building type, in the latitude and longitude coordinate system. β_2, β_2', β_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 previous models. $\beta_0^{(i)}$ represents the intercept coefficient for building structure "i" such that "i" is equal to one of number in "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 of the situation in Beijing.

Model Concerns

There are uncertainties in the authenticity of the data. Specifically, as mentioned in the Appendix, when using an influence plot to diagnose the model, it 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 using the several values of the predictor coefficients and their results on the model which they served to make predictions. I will also continue to sketch various statistical graphs 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. It 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. 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

Houses outside the area of Beijing will be regarded as meaningless negative numbers, so this also explains why the distance reaches a certain point, X times β_1 will be greater than β_0 and turns Y_1 into a negative number.

The Interaction Between LivingRoom and DrawingRoom & Price

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.

The Figure 3 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 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

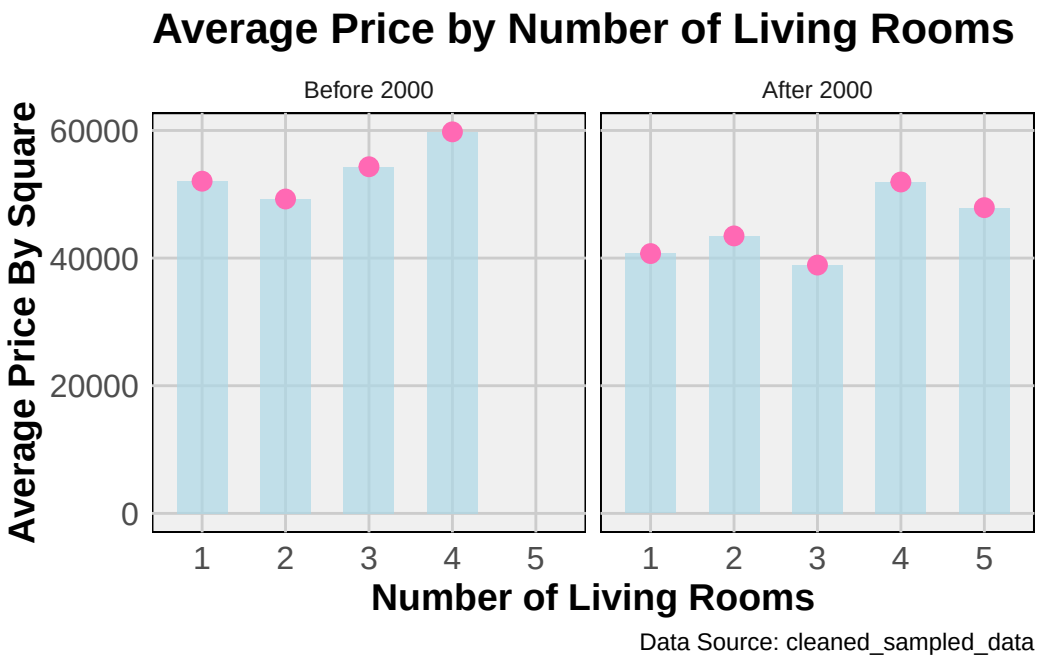


Figure 3: Prices with different number of living room

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 shows 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 4 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 drawing 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 5 living rooms in the last case, houses with three drawing rooms emerged and occupied a dominant position after 2000, with an average price per square meter of 50,000. However, the situation of one and two drawing rooms doesn't show much change after 2000.

The Interaction Between Construction Time & Price

Different construction times can also be used as potential factors to affect housing prices in Beijing. Figure 5 uses the year 2000 as the dividing line in the x-axis, and the y-axis is the price

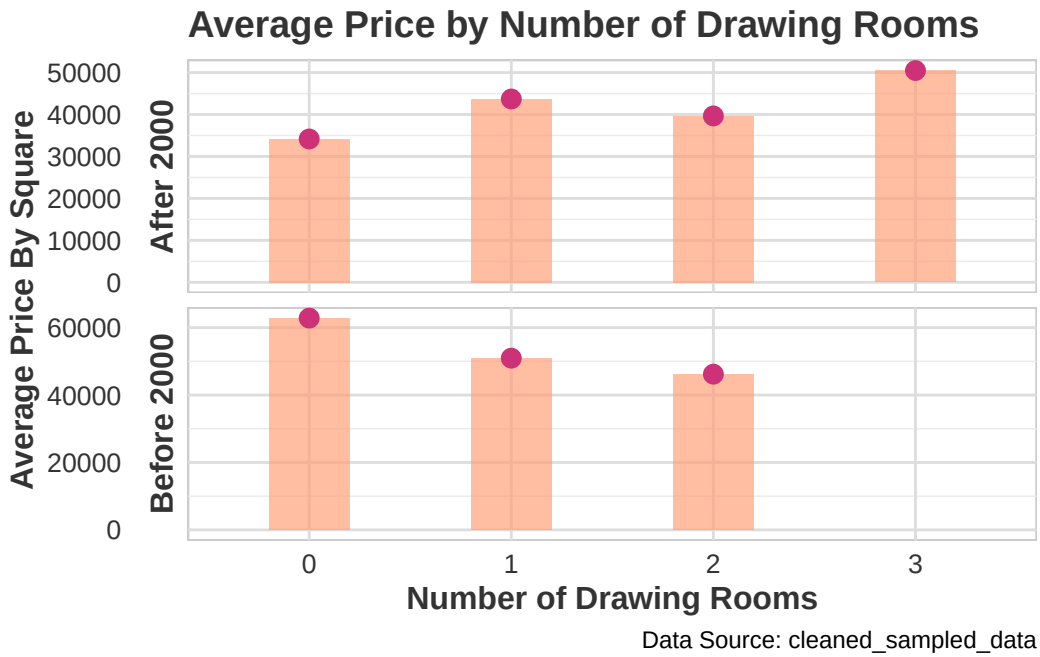


Figure 4: Prices with different number of drawing room

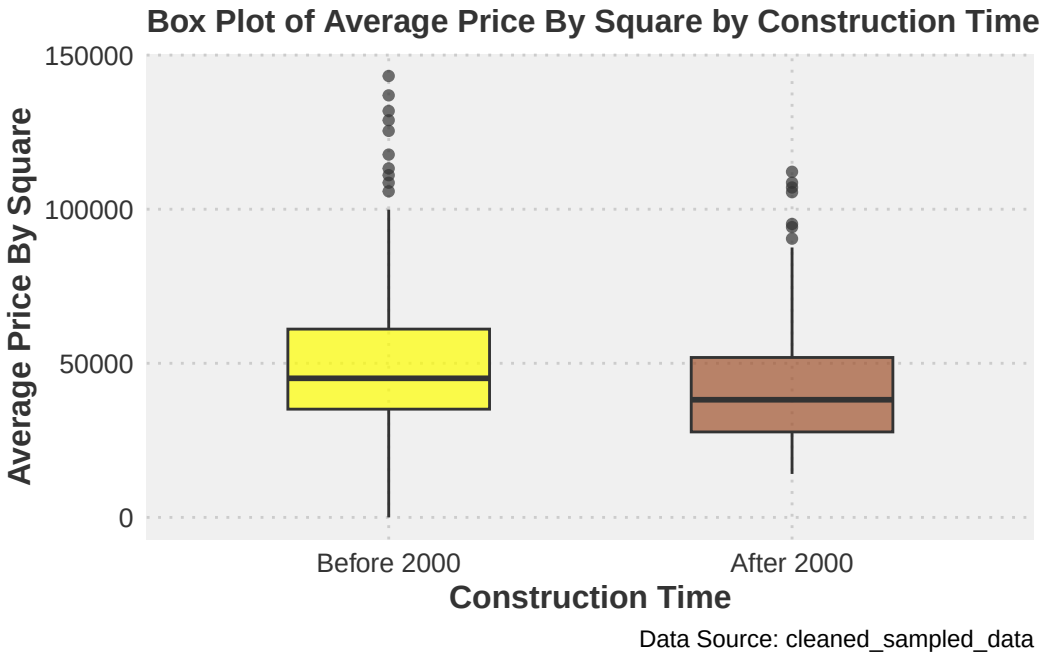


Figure 5: Prices with different construction time

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.

The Interaction Between Building Type & Price

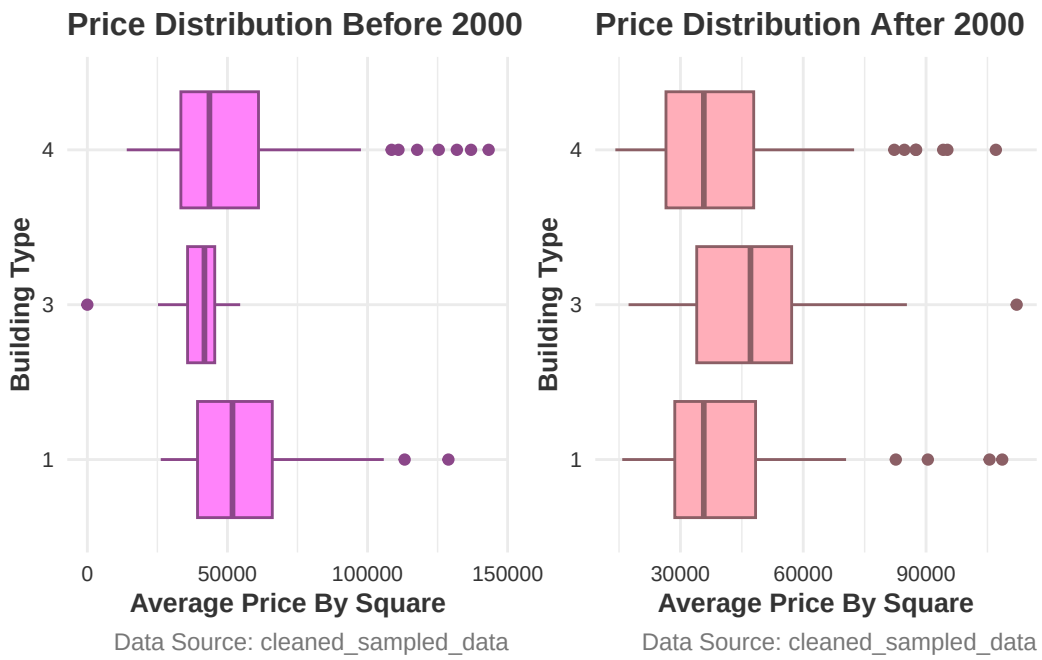


Figure 6: Prices with different building type

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 Figure 6 to view the distribution and determine the impact of differences in building types on prices. The 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 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, the median line of the two building types for plate 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%.

The Interaction Between Building Structure & Price

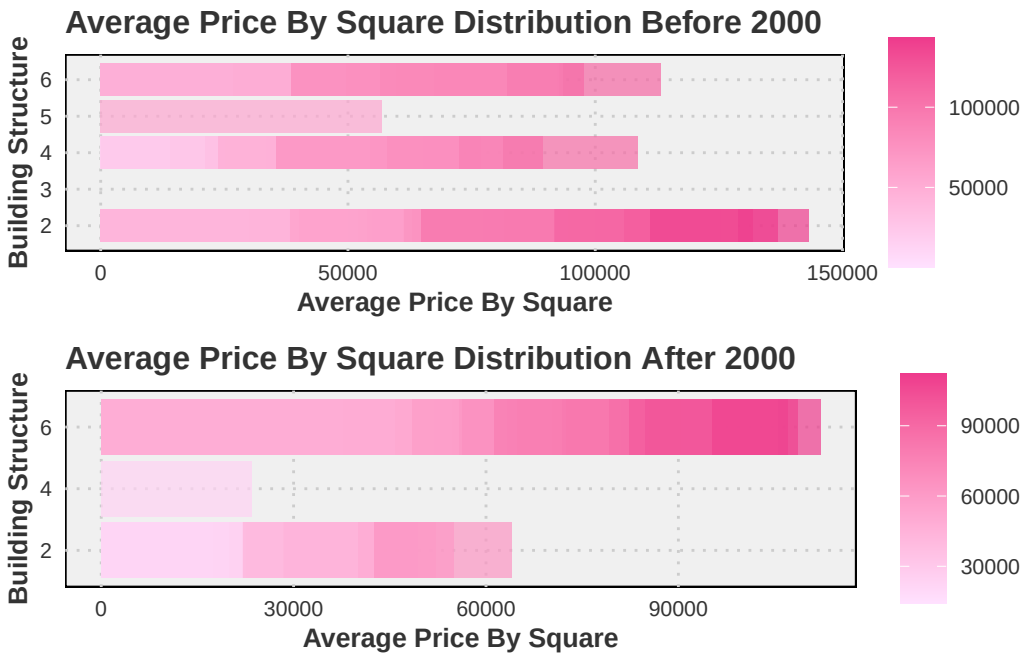


Figure 7: Prices with different building structure

Lastly, I speculate that building structure also plays a certain role in the surge in housing prices in Beijing in the 21st century. Figure 7 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 cleaned dataset, 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 dataset only included four-building structure types: mixed, brick and concrete, and 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 cheaper 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 the dataset, 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 also used to reflect the relationship between housing prices and five factors that have the potential to play decisive roles in housing prices through various types of charts. Finally, I found that the number of living rooms and building structure are also the two main factors that will affect the increase in housing prices in Beijing in the future.

Model Findings

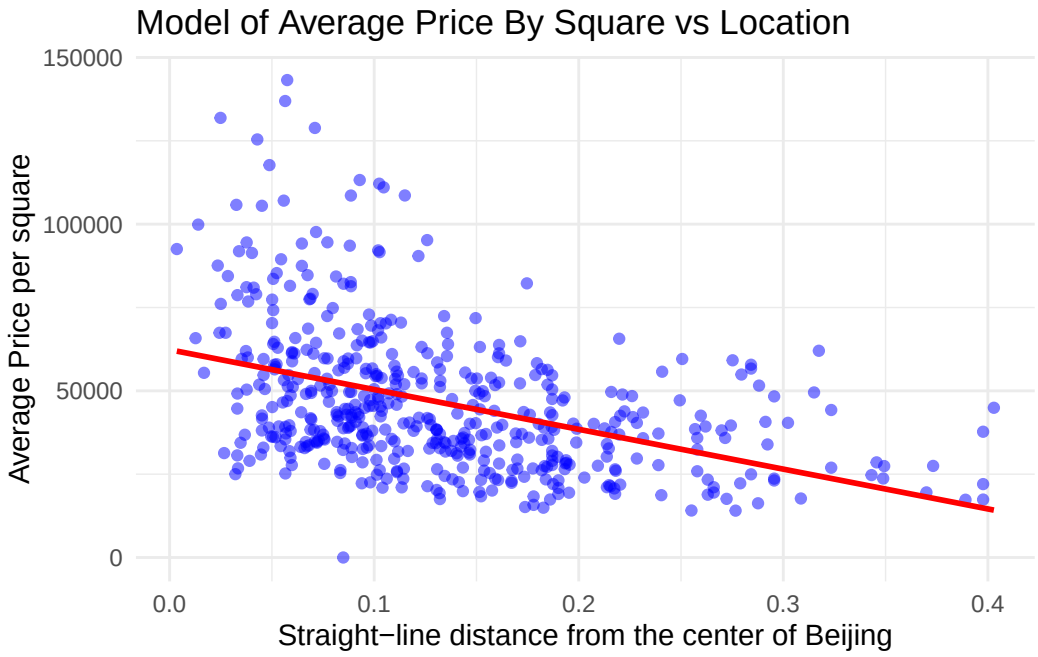


Figure 8: Prices with different location

Figure 8 shows the first model rendering linear regression. It 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 at the beginning of the Result part, which is a large number in absolute terms. This shows that for every unit (110km) farther away from the center of Beijing, 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.

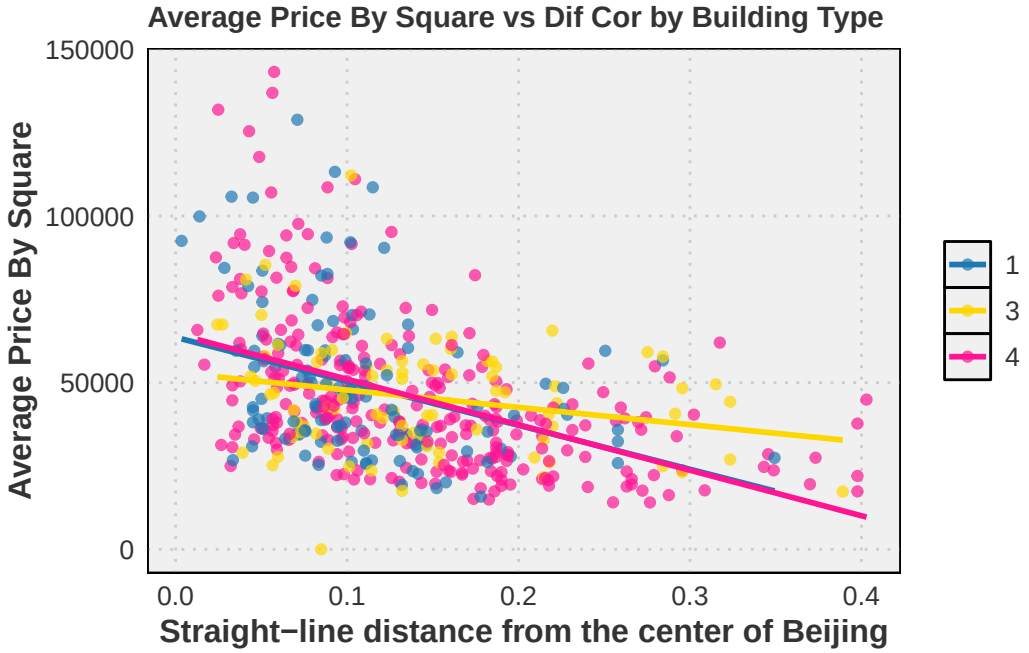


Figure 9: Prices with different building types model

For linear regression graphs Figure 9 and Figure 10 represent the second and third types of multiple linear regression respectively. For the three building types represented by the second model, including tower(1), a combination of plate and tower(3), and plate(4). Figure 9 shows that the slopes are all negative and the largest difference in 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 10 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

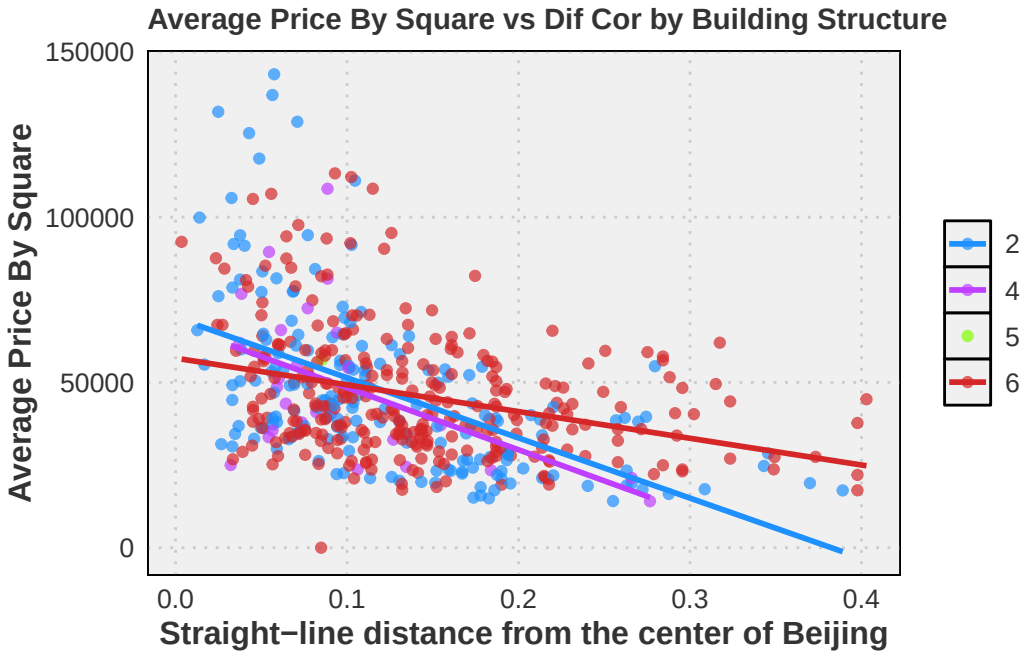


Figure 10: Prices with different building structure model

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 for each building structure by exchanging β_0, β_1, β_2, β_3 using linear regression equations. 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 in 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 drawing rooms through histograms.Figure 3 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, although the three drawing rooms only emerged after 2000, by analyzing Figure 4, the occurrence of this unprecedented situation did not have a great impact on the prices of 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 Figure 5. There is no big price difference between boxes before 2000 and after 2000, which is the middle 50% of the data. The median line around 45,000 also tells me that there isn't a huge price difference 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.

Lastly, I used a more intuitive data chart to compare the impact of building types and building structures on housing prices in Beijing. By observing the middle 50% of all building types in the Figure 6, all three types showed similar price distributions around 2000. This suggests that the surge in house prices in the 21st century was not driven by building type. Although the median line of tower-type buildings has a decrease rate of about 54%, this does not mean that the price growth is dominated by other types, because the remaining two types of median lines remain stable. On the other hand, by observing Figure 7, the depth of color represents the degree of expansiveness. It can be found that the mixed structure that dominated prices before 2000 did not have as great an impact on housing prices as the steel-concrete composite after 2000. The price reduction per square meter for mixed structures is 55%. Through horizontal comparison of results on building types and building structures, different building structures are more decisive in the growth of housing prices in Beijing.

The Role of Elevator

When sorting out the original database downloaded from Kaggle, I found that the Lianjia company also displayed information about whether each house had an elevator. This triggered my thinking about whether the elevator configuration affects the housing prices in Beijing. From a common-sense point of view, the elevator, as an effective means of transportation, can directly help some residents who are unable to walk to move between floors. Properties with elevators should have a higher average list price per square meter than properties without elevators. However, Figure 11 told me another story. Each point in this statistical chart represents an observed house, and the x-axis is used to show whether each house is equipped with an elevator, where 1 represents an elevator and 0 represents no elevator. The y-axis

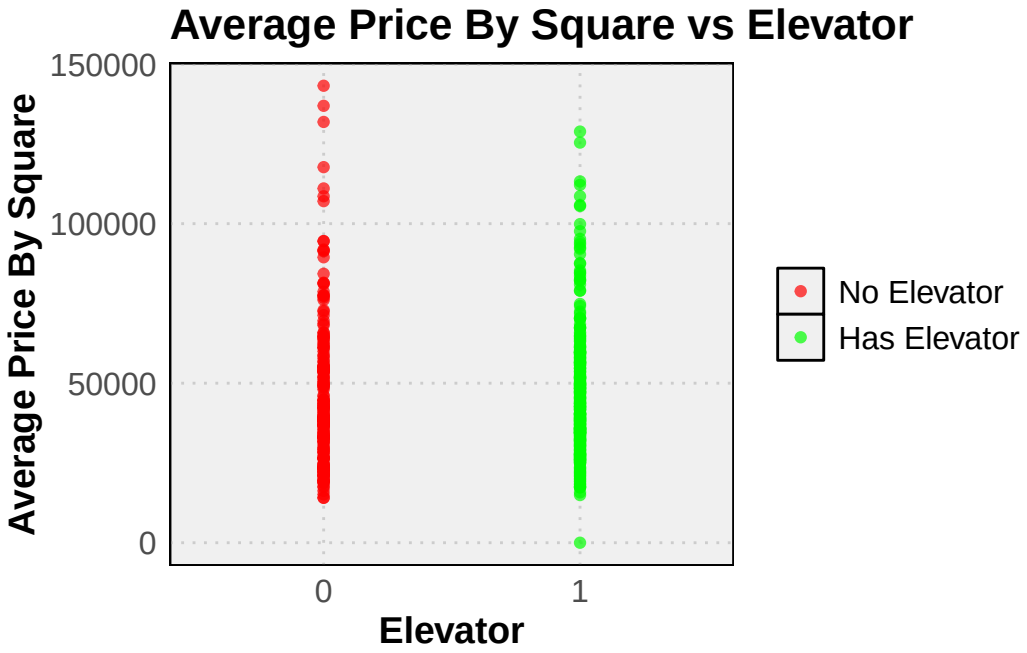


Figure 11: Prices with elevator or not

represents the price per square meter, in RMB. At prices above 100,000, the chart clearly shows that there are higher and more peak red points for buildings without elevators than for buildings with elevators in green points. This study shows that for housing prices in Beijing, the existence of elevators that make residents more convenient will not lead to higher prices.

Bias and Weakness

The source of the original data set is Lianjia Company, which is a representative real estate brokerage platform in China (X. 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 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 in the model and results.

The shortcomings of this report mainly stem from the timeliness of the original data set downloaded through the API on the open website Kaggle. Specifically, the data set used in this report comes from the data recorded by the author RUIQURM about Lianjia Company released 6 years ago. Although Lianjia Company publishes all data and sources on the open website https://bj.lianjia.com/chengjiao, the timeliness of the database only covers a total of 6 years from 2011 to 2017. Such a short period lacks credibility as a database to analyze

the changing trends in the decades from the 20th century to the 21st century. I think a suitable data set should include at least a time frame of around 15 years, such as 2000 to 2015. Fortunately, since the database used in this report has a sufficient number of listings to support the authenticity of the study, the missing and meaningless data can be ignored for the total number of observations. This also ensures that the data in the models and statistical charts established in the report are valid. In general, a longer period data set is needed as a basis for studying the influencing factors of Beijing housing prices and enhancing the accuracy of forecast trends.

Next Steps

The possible bias situation mentioned above and the problem of the short period in the original data set can be solved through some effective methods. For the former, reviewing the authenticity of each observed listing in the data set is a straightforward way to eliminate bias. The specific steps are to check, for each observation that exists in the cleaned data set, whether each given variable is consistent with what is recorded in the data set. Although this method has a relatively large time cost, it can completely remove biased data. For the latter, leveraging database merging is a preferable approach. After obtaining data sets from Lianjia Company for other periods, each separate small database can be merged into a larger data set containing larger time ranges. This will play a more convincing role in analyzing the factors affecting the surge in housing prices in Beijing and predicting housing price changes in the 21st century compared to a data set of only 6 years. Other variables that affect house prices should also be added to the data set. Because the time and angle of sunlight entering the house through the windows are different at different times, this will affect the residents' experience in the house (Ma and Narwold 2019). The orientation of the house can also be added to the data set as a key variable. Besides, whether there are landmarks such as subway stations, schools and hospitals near the house is also a decisive factor, that determines the convenience of residents' access to education and treatment (Rivas et al. 2019). Supermarkets are places that residents frequently visit is also have a significant impact on housing prices in the (L. Zhang et al. 2019). When verifying the authenticity of housing listings, these new variables can be added to the original data set to more completely determine the factors affecting Beijing housing prices.

The purpose of this report is to provide some suggestions on economic management and purchase decisions for potential and interested people who want to buy a house in Beijing by analyzing the variables provided by the data set and the price per square meter of the house. As one of the cities with the most prominent housing price growth in China, Beijing is highly representative of this research topic. Compared with the 20th century, this report focuses on three factors that determine the growth trend of Beijing's housing prices in the 21st century: geographical location, number of living rooms, and building structure. When it comes to choosing a house in Beijing, buyers can estimate the budget they need to prepare for this type of house based on their own household and house type needs, which can be used to consider whether they can save more money by giving up some unnecessary needs. In addition to

potential property buyers, economists market analysts and government policymakers will also be interested in the determinants of housing prices in Beijing. For the former, professionals can use the analysis steps and results in the report to understand broader economic trends in the field of Beijing housing prices, which will help predict the direction of real estate. For the latter, the knowledge this report brings to the world will lead to improvements in key housing policies and have a significant impact on policy guidance related to housing affordability, urban transformation, and economic progress.

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
```

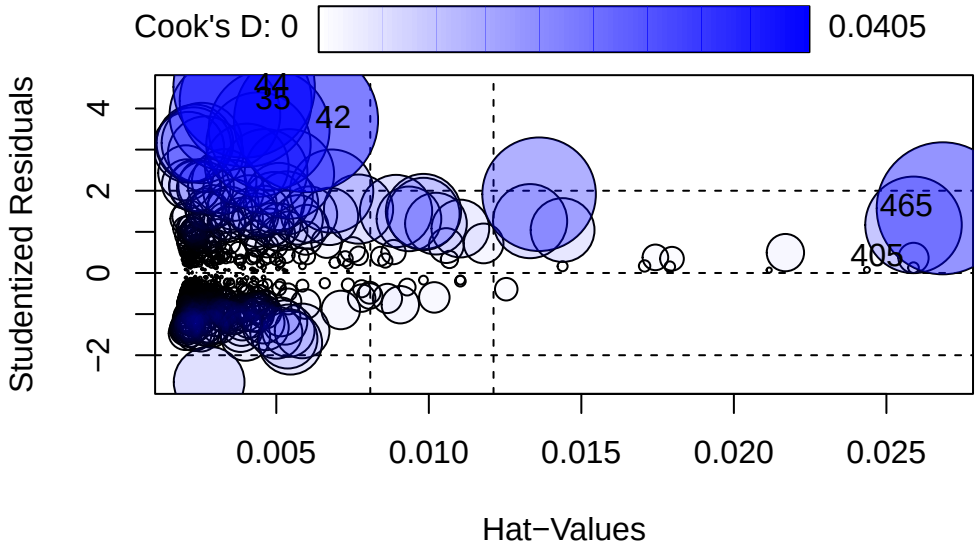


Figure 12: Influence plot for the first model

Some meaningless outliers during data cleaning can be observed from the influence plot. Through observation of Figure 12, the outliers are significantly marked in the figure. These five numbers represent the corresponding number of observation rows, indicating that they are far away from the values in this report when building the first model. A vertical line of data points is expected in the environment, and such points will have a large slope effect on the output regression linear equation when generating the model. For example, some housing prices per square meter are far beyond the norm or far below the norm. In addition, the sizes of the points corresponding to these outliers are relatively large in terms of area, which shows that these abnormal data points play a great role in the model parameters. For Figure 12, it is the result of the data set after processing the first set of outliers, just as described in the Data section. This is the same as the model concern in the Model section, if we remove the outliers of the influence plot generated by the newly generated data set from the data set every time, we will lose reliability because the number of samples does not meet the standard.

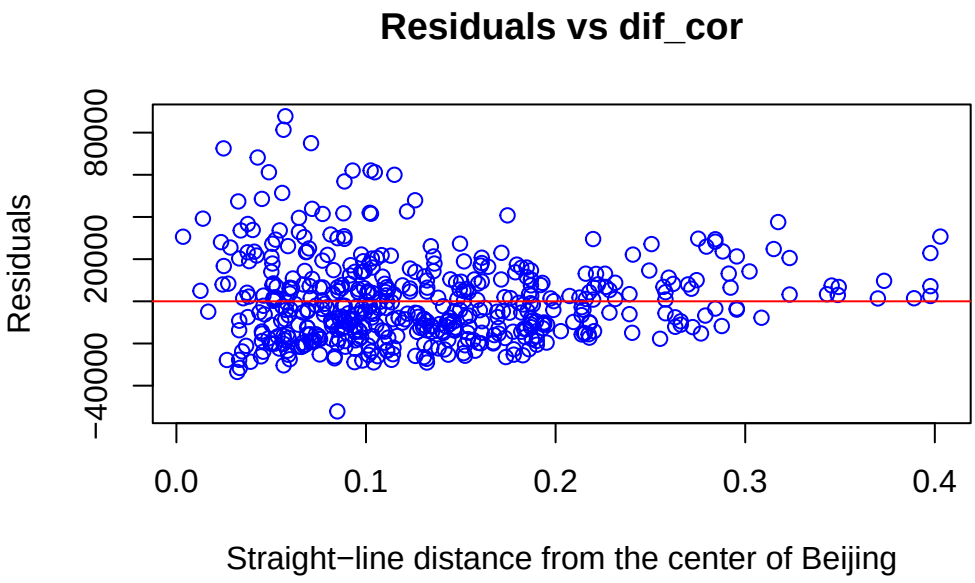


Figure 13: Residuals for the first model

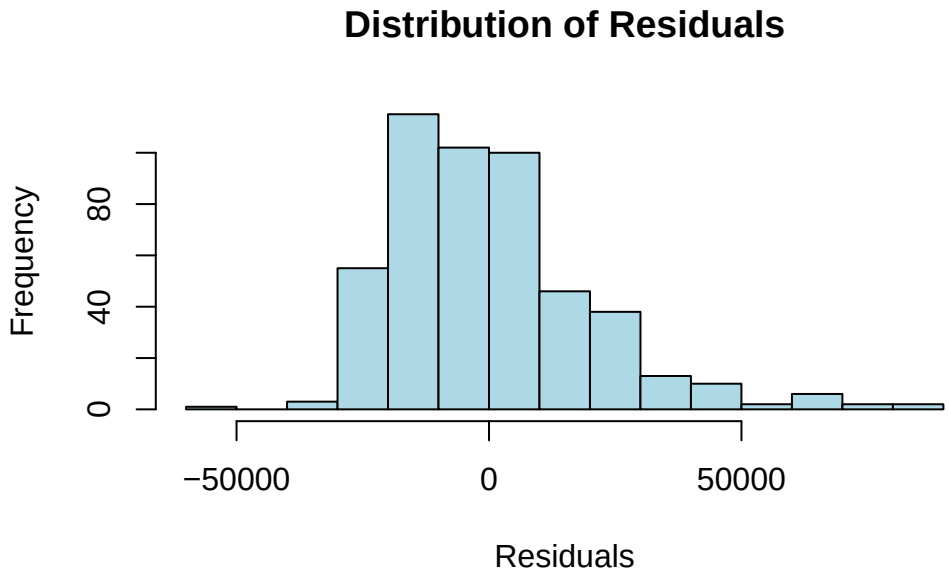


Figure 14: Distribution of residuals for the first model

Figure 13 showed me the residual plot of the first model, which consists of the straight-line distance from the center of Beijing on the x-axis and the residual represented by the y-axis. The randomness is well demonstrated in this picture, which can be judged from the fact that each dark blue hollow point in the picture is randomly scattered around the horizontal red points without showing a certain regular shape. Randomness ensures that the first model captures the underlying relationships in the data well when dealing with the observations in the data set. The distribution of residuals in different ranges can be represented by Figure 14. In the plot, the height of each bar represents the frequency of residuals falling within that bar. The widely distributed horizontal axis also indicates that the model is effective.

Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - This data set was created to record Lianjia's collection of information on housing listings within Beijing. It is mainly reflected in the summary arrangement of the location, room type, year, and other information of each property. There is no specific gap that needs to be filled.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The raw dataset was created by Lianjia company. The cleaned dataset was created by the author of this paper to serve the paper's purposes.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - No monetary cost was required in the creation of the dataset; it was created free of cost using the programming software R (R Core Team 2024a).

4. *Any other comments?*

    - Since the original database only contains data for a total of six years from 2011 to 2017 and contains some missing data and meaningless data, the reliability of the report results needs to be confirmed.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The entire original data set is composed of a table, containing 318851 house information and 26 variables including living room number, location, etc. without any photos or documents. The main research location is Beijing, China.

2. *How many instances are there in total (of each type, if appropriate)?*

   - In total, 318851 house information in rows and 26 variables in columns.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset does not contain all possible instances. First of all, I have no evidence that Lianjia put all Beijing housing information from 2011 to 2017 on a public website. Secondly, it is impossible for tens of thousands of people to buy houses every day through one company, Lianjia. Therefore, this database cannot contain all relevant data.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of 25 features. For example, the number of living rooms, the number of kitchens, the number of bathrooms, geographical location (expressed by longitude and latitude), building type, building structure, etc.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Each observation has a status id which is shown in the raw data. This status id is unique.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - There are certain missing data and meaningless data in the original data set. The reason may be a machine failure or error during recording. This may be intentionally corrupted data in order to improve company performance.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Relationships between individual instances are made explicit. The data corresponding to each variable is displayed completely on Kaggle and includes the unit.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - I recommend treating construction dates before 2000 and after 2000 separately. This is helpful in predicting the future direction and leading causes through the trend of housing prices from the 20th to the 21st century.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - There will be some unusual numbers in the data set, such as the price per square meter being unexpectedly high or low. This situation will be deleted directly to avoid affecting data analysis.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The selected data set comes from the external Kaggle, an open data set usage and review platform. I can be sure that the data set is real because I logged in to https://bj.lianjia.com/chengjiao and successfully obtained the downloaded data set. There are no restrictions on downloading the dataset.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The dataset may contain confidential data if the residents label their house trade as private.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - The dataset not contain offensive, insulting, and malicious text.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset does not identify any sub-populations.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - It would not be possible to identify individuals from the dataset.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The data not contains the data that might be considered sensitive.

16. *Any other comments?*

    - None

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.* -The data is obtained by Lianjia's data loggers by observing the characteristics of each listing. There is no data derivation process, everything is directly observed.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was collected using beta API from Kaggle, which can be found in the setting.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - This data was not obtained from a larger database. The Sampling strategy is to sort and collect samples according to the construction time of the building.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Only the buyer of each house arranged by Lianjia company.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The time span of this data set is from 2011 to 2017. A total of six years. I can't judge whether the timeframe matches the creation timeframe of the data.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - None

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected from other sources, using the Kaggle API.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - The individuals in question were not notified about the data collection.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - The individuals in question consented to the collection and use of their data in the Lianjia company. The link can be found here: https://bj.lianjia.com/chengjiao.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - I don't have a way to get this information.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - None

12. *Any other comments?*

    - None

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.* -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.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - The "raw" data is contained in the repository.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - The software used to clean the data consists of R packages(R Core Team 2024a).
4. *Any other comments?*
    - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Because the original data set is publicly available on Kaggle for users to download and analyze for free, all the original data is likely to have been conducted by some other users for related research. However the cleaned dataset is only used by me in this report.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - Yes. Code and data are available at: https://github.com/zxc0707/Beijing-housing-price
3. *What (other) tasks could the dataset be used for?*

- This data set can also be used as a survey on the interest level of some Beijing house buyers. This is because the original data contains the variable followers, which describes the number of people following the transaction.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - None

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset should not be used for any malicious or illegal activity.

6. *Any other comments?*

    - None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - This data set may be used as a reference by some housing price analysts or policy makers in the government.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset will be distributed on GitHub via the repository.

3. *When will the dataset be distributed?*

    - The dataset will be distributed on April 18, 2024.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - None

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - None
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - None
7. *Any other comments?*
    - None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be hosted by the author of the study.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The owner of the dataset can be reached via the GitHub account the repository is hosted on.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - None
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - The dataset will not be updated in the future.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - There are no limits.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

- Older versions of the dataset may be hosted by the author of the study on their local files. The obsolescence will be communicated from the repository the study is hosted on.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Others can contribute to the dataset via GitHub's built-in collaboration features, which will be verified and finalized by the author.

8. *Any other comments?*

    - None

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

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.

Ma, Alyson, and Andrew Narwold. 2019. "Which Way Is up? Orientation and Residential Property Values." *Journal of Sustainable Real Estate* 11 (1): 40–59.

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/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Rivas, Ryan, Dinesh Patil, Vagelis Hristidis, Joseph R Barr, and Narayanan Srinivasan. 2019.

"The Impact of Colleges and Hospitals to Local Real Estate Markets." *Journal of Big Data* 6 (1): 1–24.

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, Ling, Jiantao Zhou, Eddie CM Hui, and Haizhen Wen. 2019. "The Effects of a Shopping Mall on Housing Prices: A Case Study in Hangzhou." *International Journal of Strategic Property Management* 23 (1): 65–80.

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.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.