

# Analyzing the Factors that Determine Housing Prices in Beijing and Predict Trends in the 21st Century\*

Xincheng Zhang

April 2, 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

<b>Introduction</b>	<b>1</b>
<b>Data</b>	<b>3</b>
Dataset . . . . .	3
Variables and Features . . . . .	3
Missing Data . . . . .	6
<b>Model</b>	<b>7</b>
Features . . . . .	11
Model Concerns . . . . .	11
<b>Results</b>	<b>11</b>
LivingRoom and DrawingRoom & Price Interaction . . . . .	11

---

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

Construction Time & Price Interaction . . . . .	14
Building Type & Price Interaction . . . . .	14
Building Structure & Price Interaction . . . . .	14
Elevator & Price Interaction . . . . .	14
<b>Discussion</b>	<b>14</b>
Data and Model Findings . . . . .	14
Weaknesses and next steps . . . . .	14
<b>References</b>	<b>16</b>

## 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
price	The average price by square
square	The square of house
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 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)

Table 2: Details of Several Data Features

Feature	Details
---------	---------

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



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).

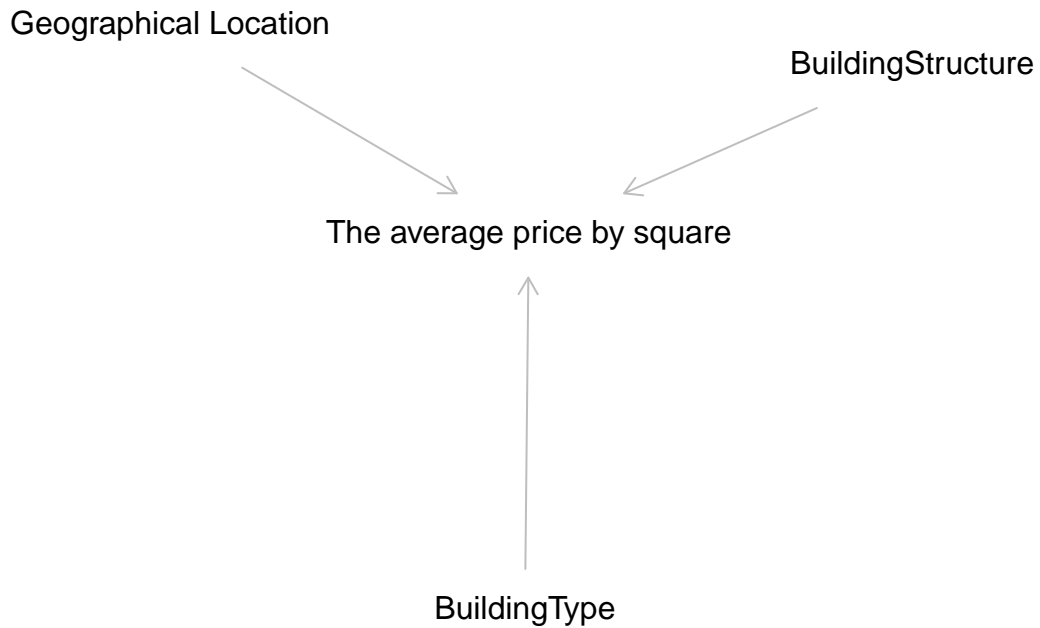
## Missing Data

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.

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 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

## Model

For this report, I tried to use a linear regression model to confirm that the location of the house, building Type, and building Structure have a linear relationship with the average price by square of the room and construct directed acyclic graph to visualize what I want to discuss and Modeled variables. This helps to clearly show the relationship I think exists between these variables.



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 ***

---

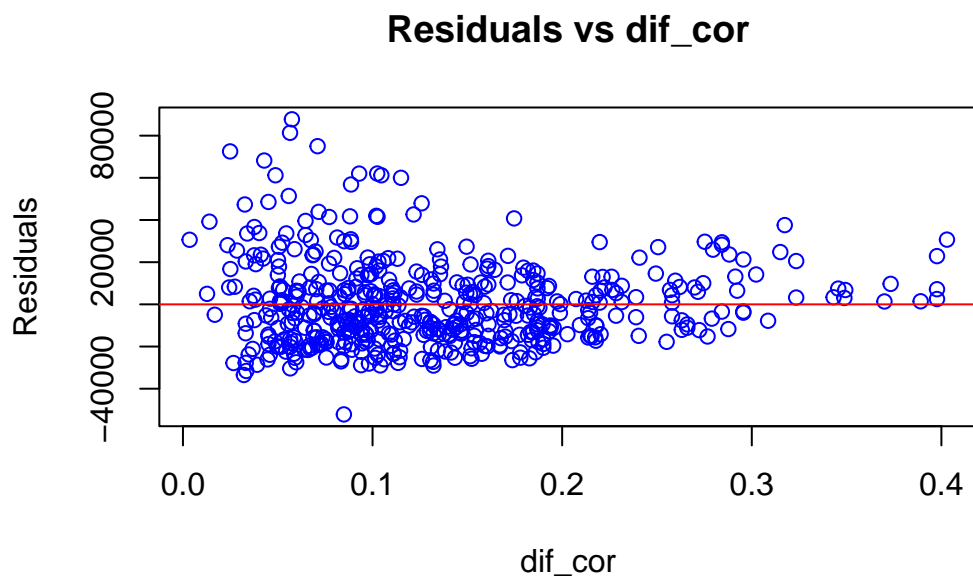
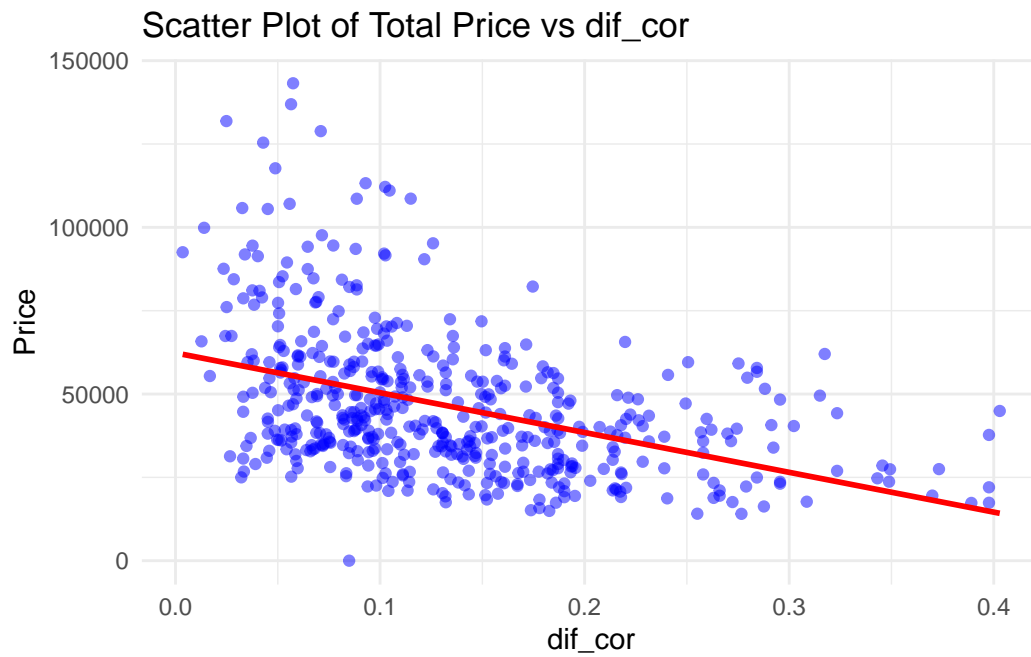
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

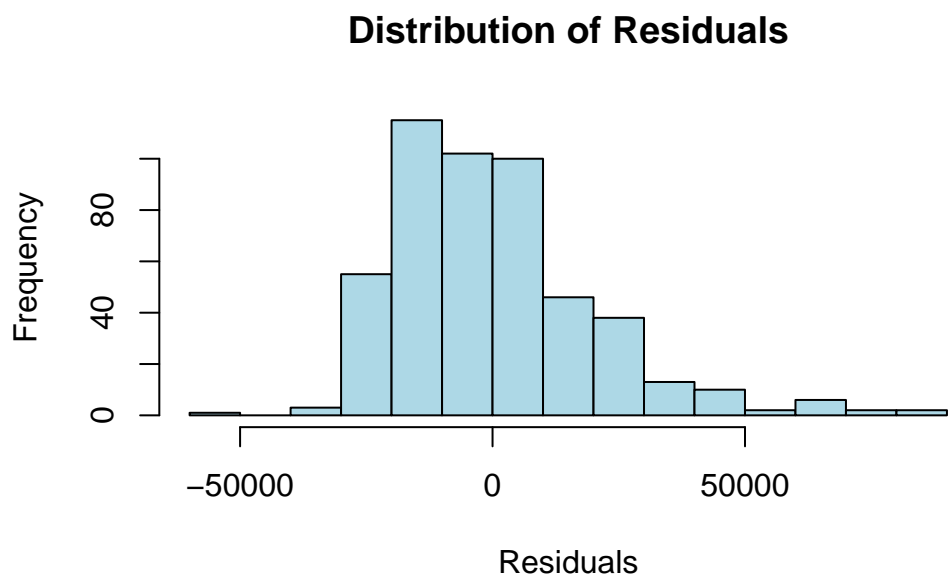
Residual standard error: 19820 on 493 degrees of freedom

Multiple R-squared: 0.178, Adjusted R-squared: 0.1763

F-statistic: 106.7 on 1 and 493 DF, p-value: < 2.2e-16







$$Y_1 = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

X means dif\_cor

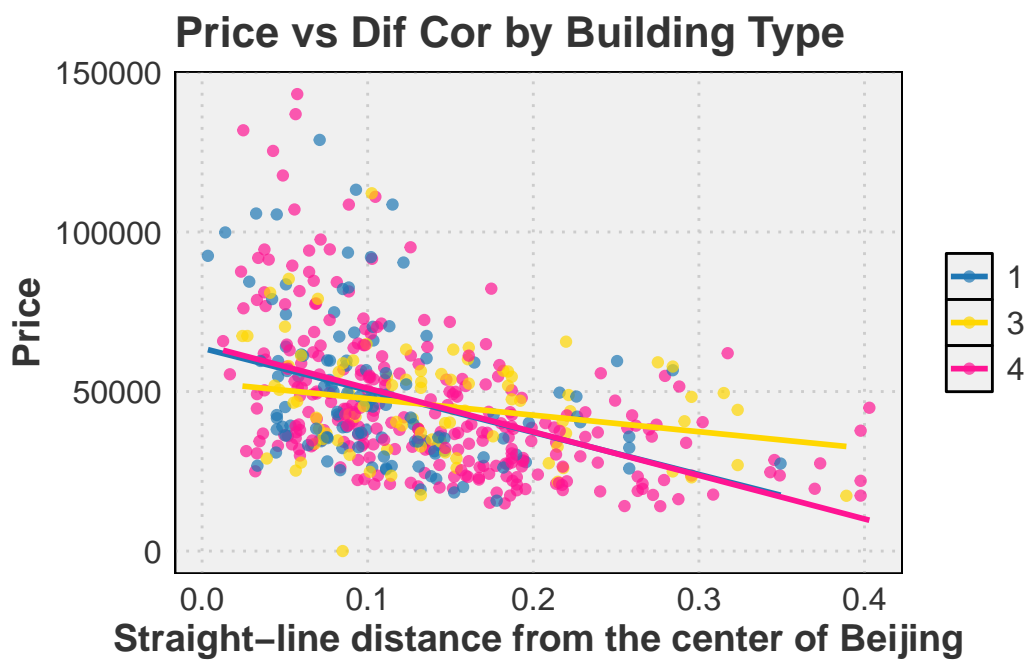


Figure 2: Prices with Different Building types

$$Y2 = \beta_0 + \beta_1 X + \beta_2 L + \beta_3 D + \epsilon \quad (2)$$

X means dif\_cor

L means number of living room

D means number of drawing room

$$Y3 = \beta'_0 + \beta'_1 X + \beta'_2 L + \beta'_3 D + \epsilon' \quad (3)$$

$$Y4 = \beta''_0 + \beta''_1 X + \beta''_2 L + \beta''_3 D + \epsilon'' \quad (4)$$

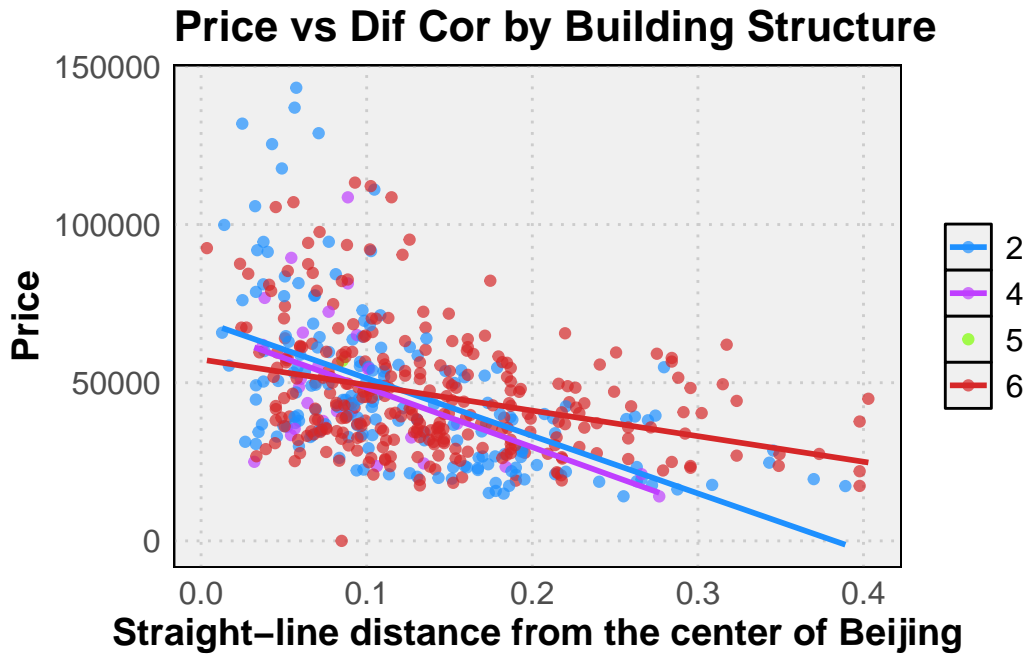


Figure 3: Prices with Different Building Structure

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

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

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

Features

Model Concerns

Results

LivingRoom and DrawingRoom & Price Interaction

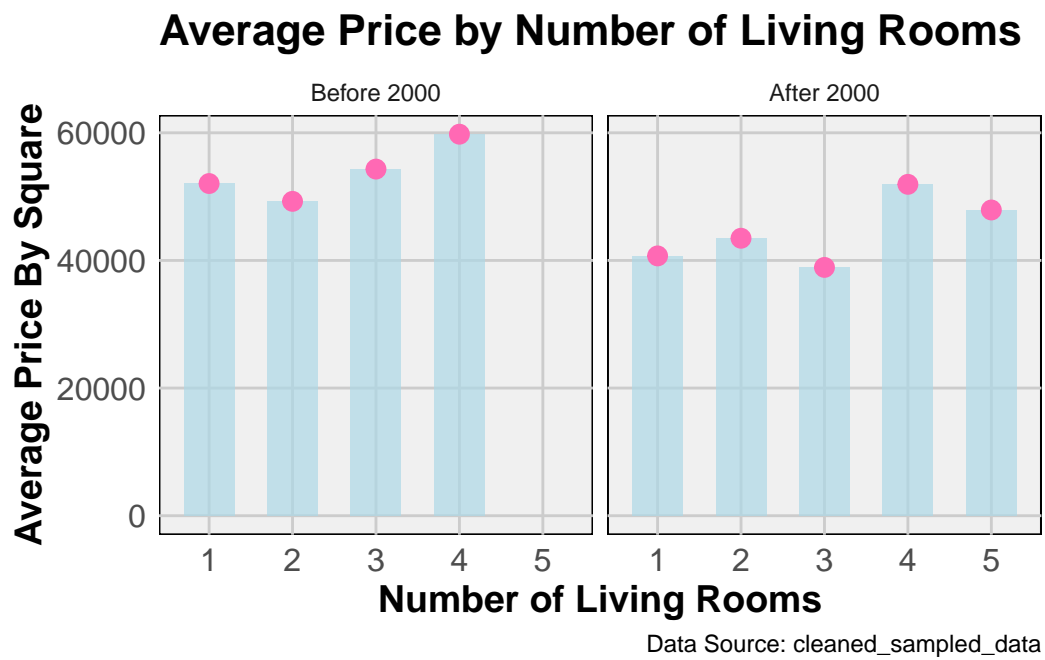


Figure 4

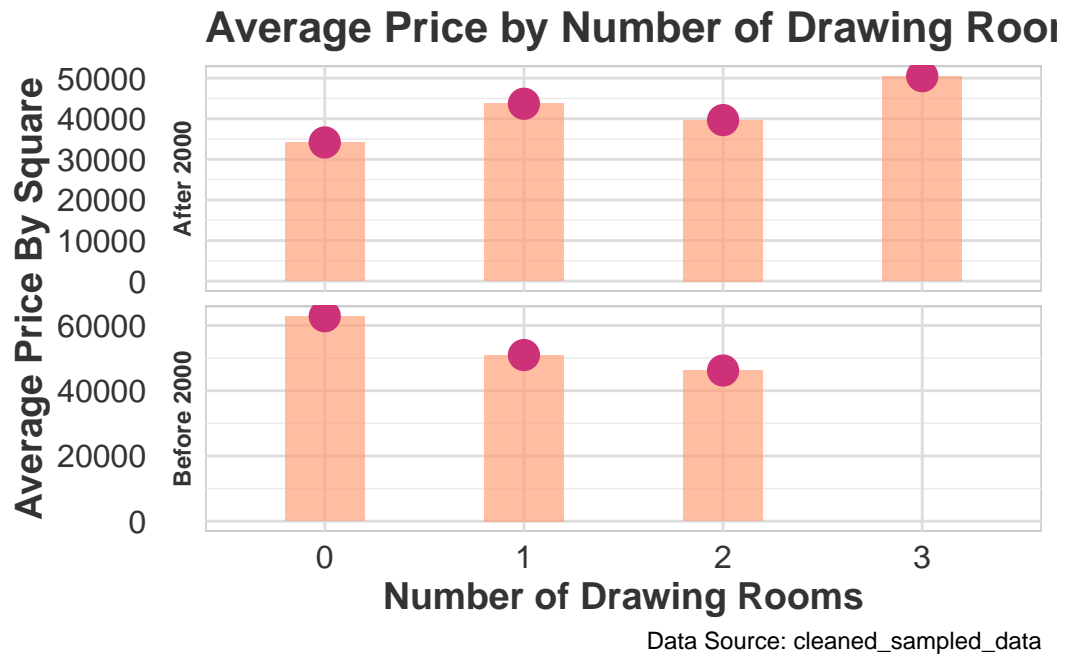


Figure 5

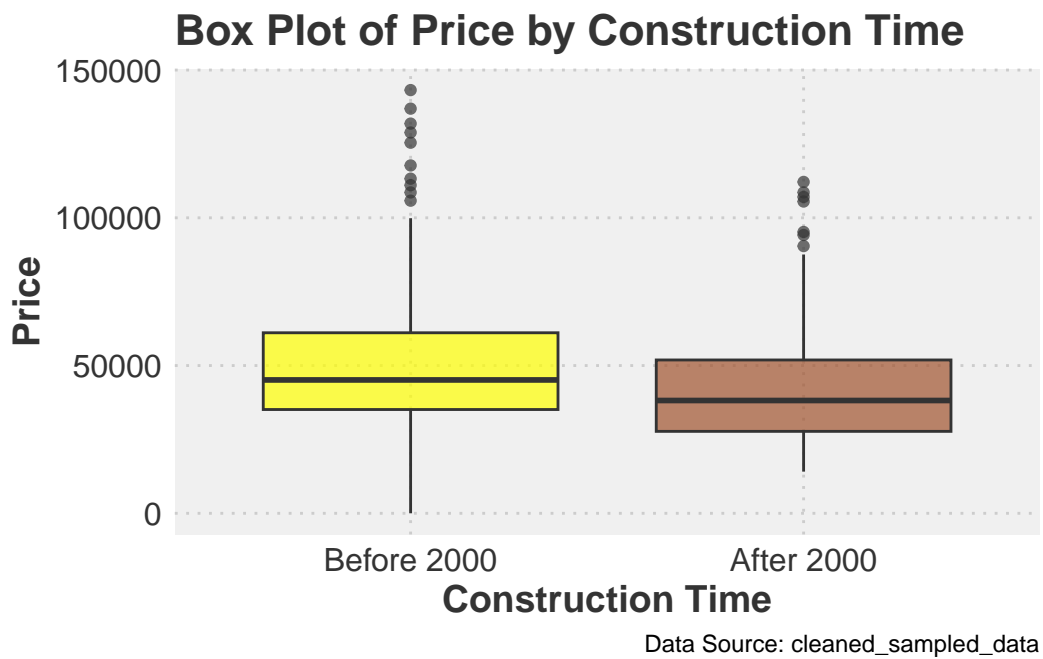


Figure 6

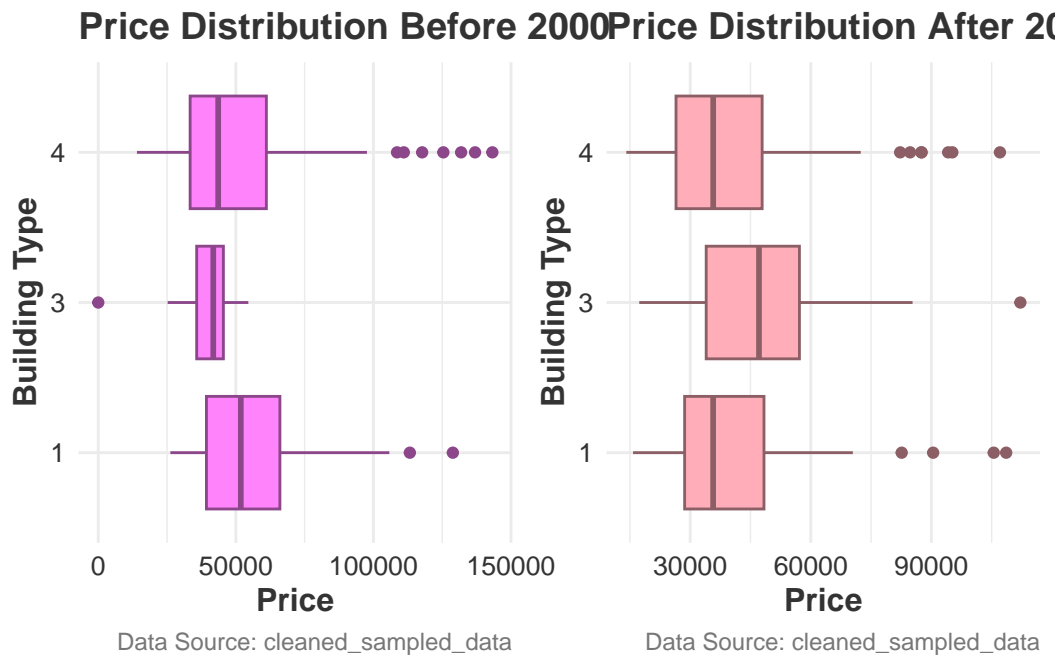


Figure 7

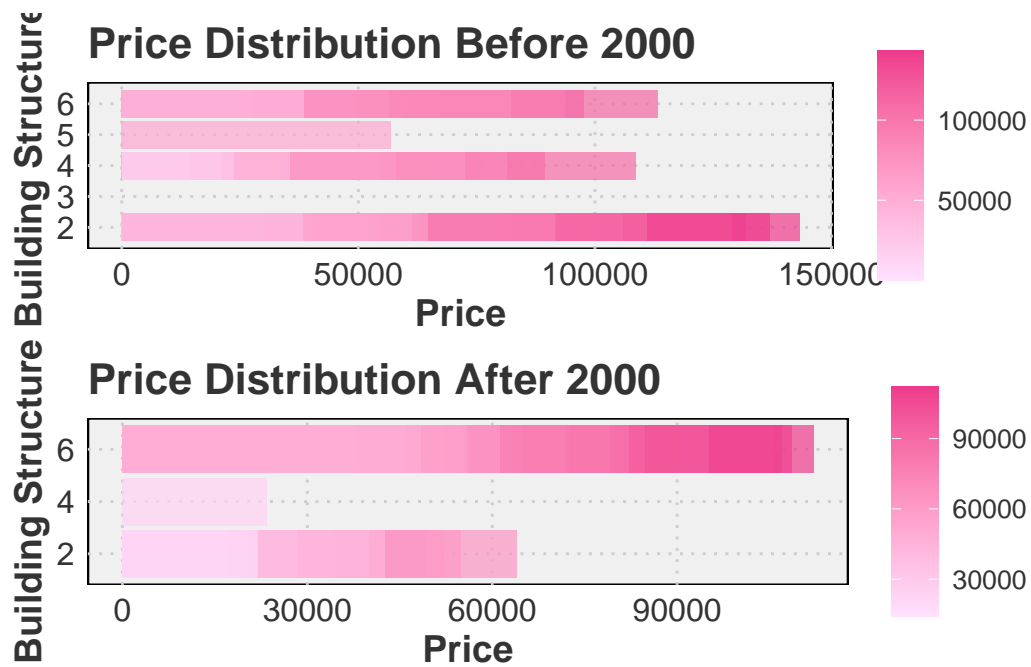


Figure 8

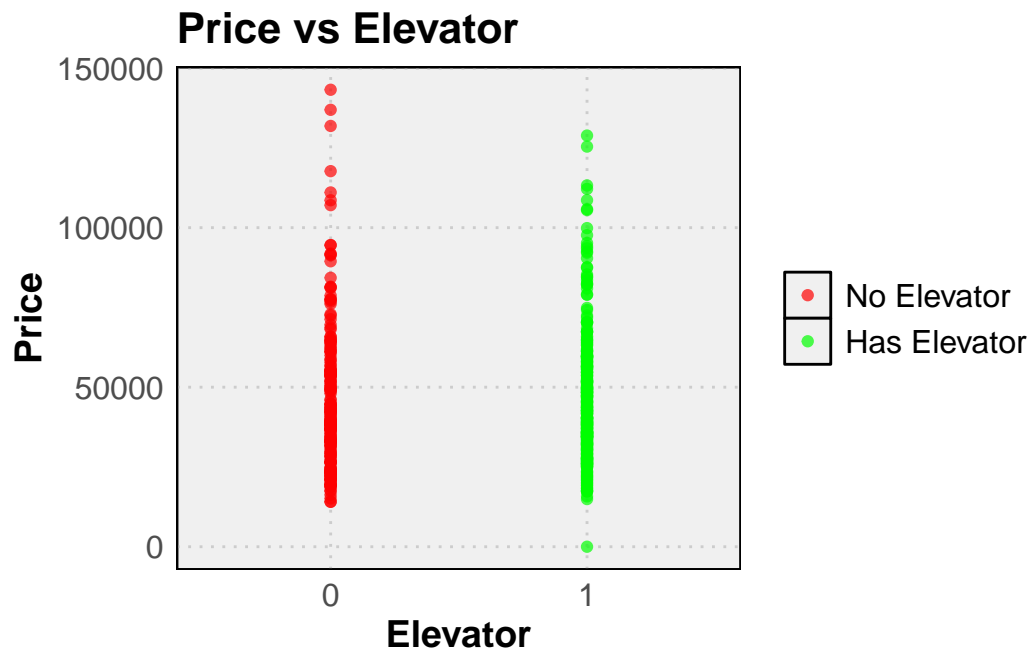


Figure 9

**Construction Time & Price Interaction**

**Building Type & Price Interaction**

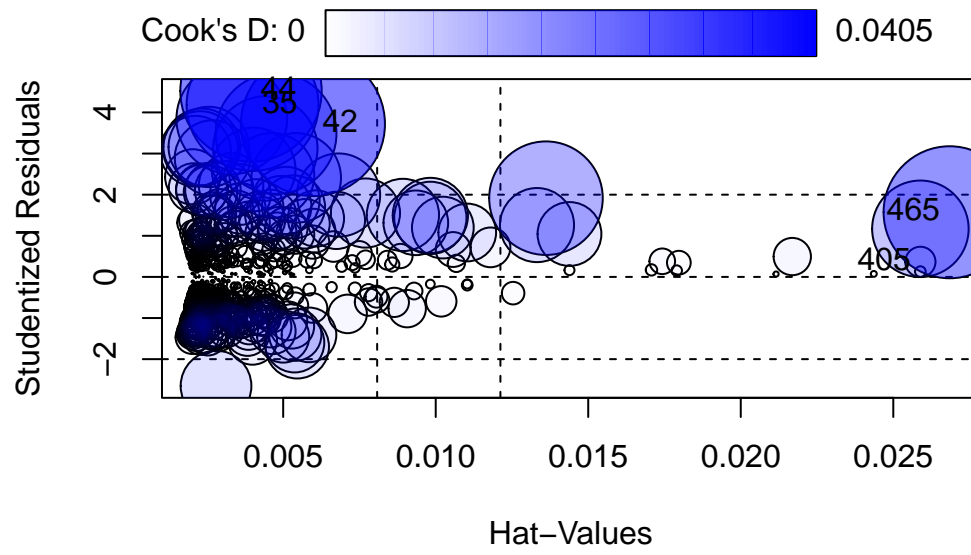
**Building Structure & Price Interaction**

**Elevator & Price Interaction**

**Discussion**

**Data and Model Findings**

**Weaknesses and next steps**



	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



## References

- 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.
- 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.
- 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.
- 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.