An analysis of house prices in Beijing using linear regression model*

Yiyi Yao

April 11, 2024

This study investigates the influence of property size, location, and type on housing prices in Beijing over the past decade. Employing linear regression analysis on a comprehensive dataset of property transactions from 2011 to 2017, the research highlights the growing importance of technological integration alongside traditional market drivers like size and location. This paper provides crucial insights for stakeholders looking to navigate the evolving dynamics of urban real estate markets, offering a basis for future policy and development strategies in the sector.

1 Introduction

Beijing, as the capital of China, is an international metropolis integrating economy and politics, trends and culture. In recent years, the difficulty of buying a home has become a major problem for young Chinese people (Yan, Feng, and Bao (2010)). And Beijing's housing prices have been high and rising for many years (Zhang and Yi (2018)). I have selected a comprehensive set of data containing ten years' worth of housing transactions in Beijing, in the hope that it will provide an in-depth study of the pattern of Beijing's housing prices. This study examines the impact of various factors such as real estate size, location, and technological amenities on Beijing's house prices, filling a significant gap in current research that tends to ignore the impact of modern technological integration and changing demographic trends on real estate values.

This paper adopts a linear regression analysis method to study the impact of different variables on house prices using a comprehensive dataset of ten years of real estate transactions in Beijing (Liu and Ma (2021)). The methodology takes into account not only traditional factors such as size and location, but also new influences such as housing configuration, aiming to paint a

*Code and data are available at: https://github.com/Yaoee111/house-price.git. DOI available at: DOI 10.17605/OSF.IO/JS83Y.

detailed picture of the drivers of current market trends. The results of the study show that the market has changed significantly, with technological amenities becoming an important factor in determining property values in addition to traditional factors.

My estimand is finding effects of property size, location, and type of building on the total price of housing in Beijing. Specifically, we aim to fit a linear regression model to predict the house prices.

The rest of the paper is organized as follows: The Section 2 describes the data, variables, and methods used in the study and the rationale for the selection of this dataset. The data is presented through graphs. The Section 3 describes how to fit a linear regression model. The Section 4 presents the results of the analysis. The Section 5 provides an in-depth discussion of our findings and reflections on the research process. Finally, the Section A adds details of the models.

2 Data

2.1 Overview

This dataset, sourced from Kaggle, provides a comprehensive collection of housing price records in Beijing from 2011 to 2017 (with some transactions recorded as early as 2009 and as late as January 2018). This data was primarily fetching from Lianjia, which is a Chinese house trading platform.

2.2 Data Management

I use the statistical programming language `R` (R Core Team 2023). In the data analysis and visualization process, I also made use of the following `R` packages: `readr` (Wickham et al. 2024), `dplyr` (Wickham et al. 2014), `knitr` (Xie 2018), `kableExtra` (Zhu et al. 2019), `ggplot2` (Wilkinson 2011), `stringr` (Wickham and Wickham 2019), `broom` (Robinson 2014), `tibble` (Müller and Wickham 2023), `stats` (R Core Team 2013).

2.3 Variable

This dataset contains detailed variables for evaluating each property. I will focus on some most vital variables. Details are in (Table 1)

Table 1: Key Variables in the Beijing Housing Price Dataset

Variable	Description
tradeTime	The time of transaction

DOM	Active days on market
followers	The number of people following the transaction
totalPrice	The total price
price	The average price by square
square	The square of house
livingRoom	The number of living rooms
drawingRoom	The number of drawing rooms
kitchen	The number of kitchens
bathroom	The number of bathrooms
floor	The height of the house
buildingType	Including tower (1), bungalow (2), combination of plate and tower (3), plate (4)
constructionTime	The time of construction
renovationCondition	Including other (1), rough (2), simplicity (3), hardcover (4)
buildingStructure	Including unknown (1), mixed (2), brick and wood (3), brick and concrete (4), steel (5), steel-concrete composite (6)
ladderRatio	The proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident have on average.
elevator	Have (1) or not have elevator (0)
fiveYearsProperty	If the owner has the property for less than 5 years

2.4 Rationale for Selection

The selection of this dataset for modeling house prices in Beijing is based on several compelling reasons:

- This dataset contains a broad range of critical determinants of housing prices, including location, size (square meters), and the number of rooms (living, drawing, kitchen, bathroom). These variables are well-established and easy to read.
- With trade data spanning from 2011 to 2017 and some records from 2018, this dataset offers a historical perspective on the market, allowing for the analysis of trends over time.
- The dataset is sourced from Lianjia, which is one of the biggest platform in Chinese house trading market. Thus, we can trust in the quality and reliability of these data.

2.5 Limitation

While this dataset provides valuable information for modeling, it also has limitations that could impact the analysis:

- It exists missing or incomplete records, such as days on market (DOM) or construction time, which could reduce the dataset's overall utility and may introduce bias.

- As the dataset covers transactions up to 2017 (with some from 2018), the data might not reflect current market conditions, which are influenced by recent economic and policy changes.

2.6 Data cleaning and preparation

In preparing the Beijing house prices dataset for analysis, I first cleaned the data to enhance data integrity and usability. We started by stripping non-numeric characters from variables such as 'floor', where only the numerical floor levels were retained. Further, I filtered out rows with invalid entries in key columns like 'buildingType' and 'constructionTime'. The dataset then underwent general cleaning procedures, including the removal of missing values and duplicate records. This process aimed to mitigate any potential biases and provide a foundation for the predictive modeling of Beijing's house prices. Also, since the original dataset was too big (318852 rows of data in total), I randomly chose 300 rows for further analysis.

2.7 Plots

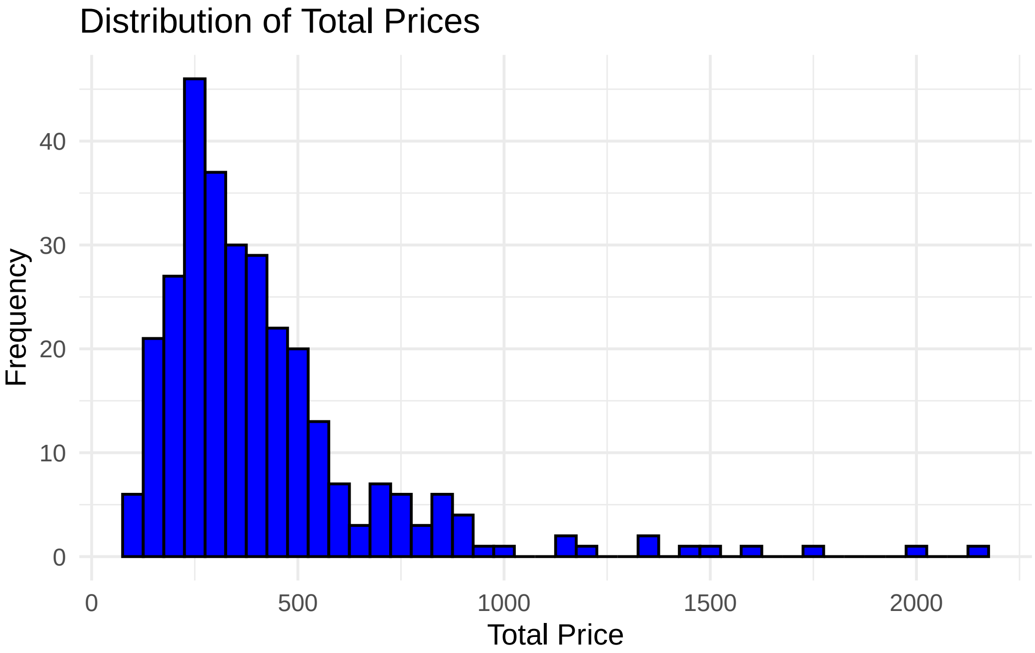


Figure 1: Relationship between Total Price and Frequency

Figure 1 shows the distribution of total prices across all properties listed in the dataset. The histogram's bins are set at intervals of 50 to effectively capture the range and frequency of

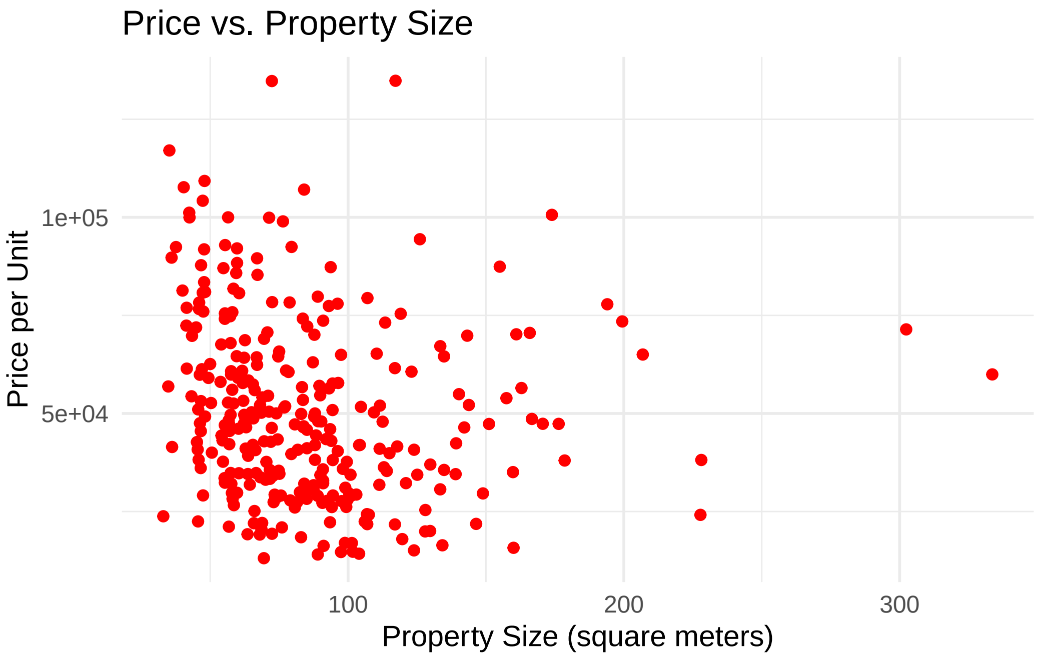


Figure 2: Price vs. Property Size

pricing. Such a plot is useful for identifying the most common price ranges and observing any skewness or outliers in market prices.

Figure 2 illustrates the relationship between the size of properties (in square meters) and their unit prices. Each point represents a property, with its position indicating both its size and its price per square meter. This visualization is crucial for analyzing how price per unit area varies with property size, potentially revealing trends like economies of scale or premium pricing for larger properties.

In Figure 3, each point represents a property, plotted according to its district and the price per square meter. This plot allows for the comparison of price levels across different districts, highlighting variations that might be due to location desirability, availability of amenities, and other socio-economic factors.

3 Model

The goal of my modelling is twofold. Firstly, I want to Explore the relevance of various variables to house prices. Secondly, I want to fit a linear regression model to predict the house prices, which involves two main phases: feature selection and model development.

Background details are included in Appendix A.1.

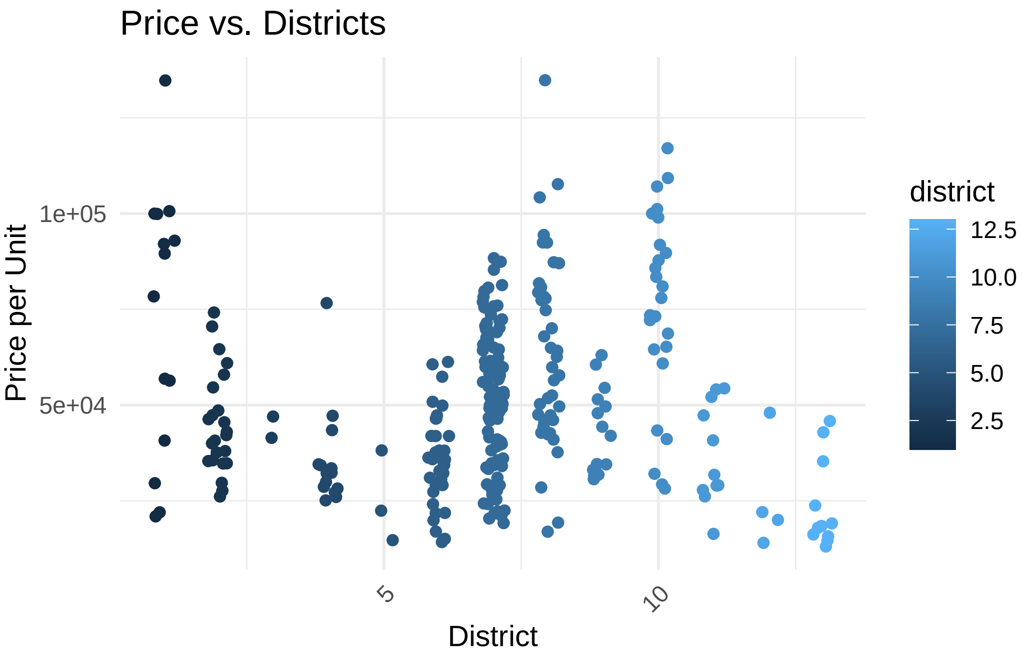


Figure 3: Price vs. Districts

3.1 Feature selection

In the context of predicting house prices, the relevance of variables can be assessed through exploratory data analysis (EDA), which includes looking at correlation coefficients between potential predictor variables and the house price (the target variable). I would examine scatter plots for continuous variables and box plots for categorical variables to visualize relationships. Statistical tests help in quantifying the strength of associations. Variables with strong correlations or significant test results could be considered relevant.

3.2 Model set-up

3.2.1 Linear regression model

To predict house prices, I will fit a linear regression model. The general form of a linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where: - Y is the target variable (house price). - X_1, X_2, ... , X_n are the predictor variables (such as square meters, number of rooms). - β_0 is the intercept. - β_1, β_2, ... , β_n are the coefficients representing the weight of each predictor. - ϵ is the error term, representing the part of Y the model cannot explain.

Each coefficient indicates how much the target variable is expected to increase when that predictor variable increases by one unit, holding all other predictors constant.

3.3 Model justification

Choosing between a linear regression model and a Bayesian model for predicting house prices in Beijing depends on their distinct advantages. A linear regression model is simple and ease for interpretation, and compute in high efficiency. Therefore, it is ideal for initial explorations and for direct understand on how different variables affect house prices. In terms of results, we can expect linear regression models to provide clear, straightforward estimates of the impact of each factor on house prices.

4 Results

4.1 visualize relationship between variables

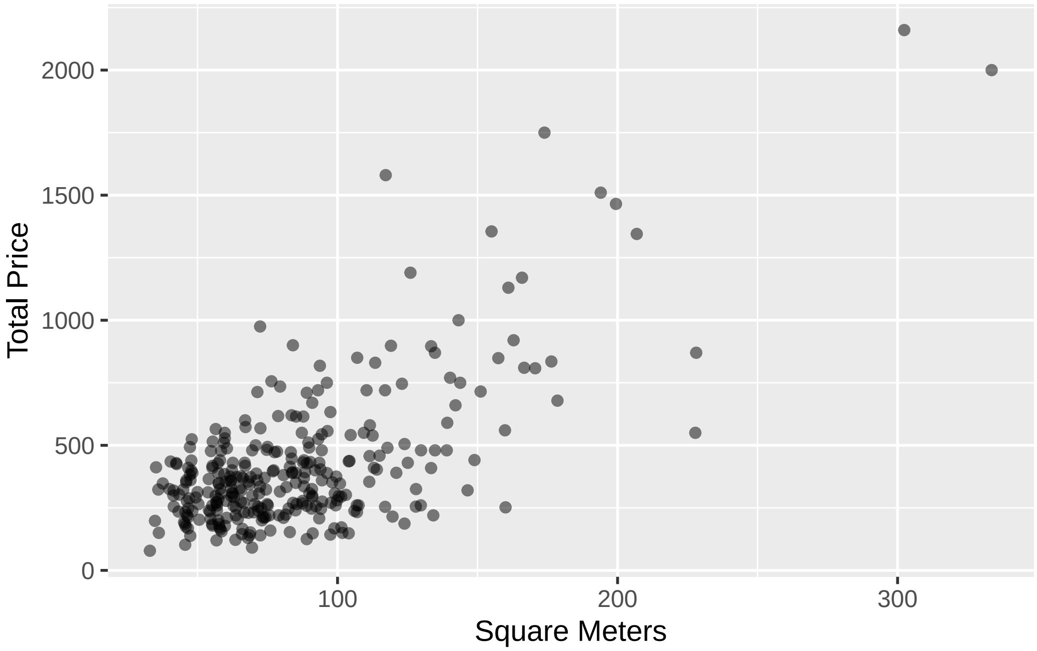


Figure 4: Total Price v.s. Square Meters

From the scatter plots (Figure 4), we look for trends indicating relationships between continuous variables (totalPrice) and square (the size of the property in square meters). We can find a positive correlation, which appear as an upward trend in the scatter plot. The shape of the plot suggests that as the size of the property increases, so does its price.

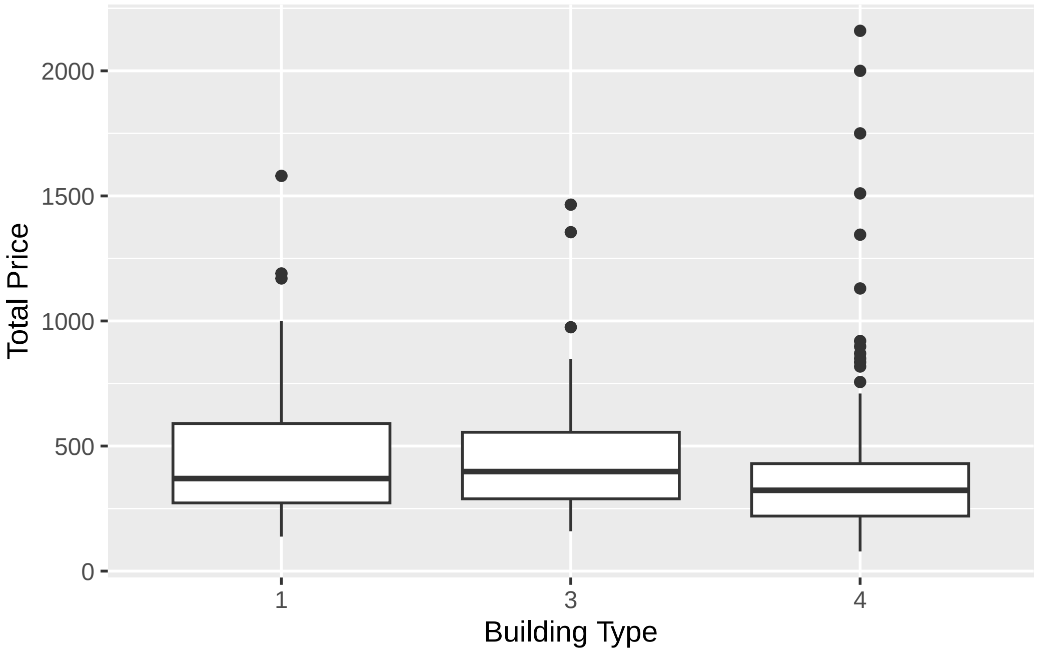


Figure 5: Total Price by Building Type

The box plot (Figure 5) for categorical variables, like buildingType, provides insights into how the prices are distributed across different types of buildings. For example, differences in median prices, the presence of outliers, and the spread of prices within each category can reveal which types of buildings tend to be more expensive or show greater variability in prices.

4.2 Visualize linear regression model

Figure 6 clearly visualizes the relationship between property size and total price in the Beijing housing market. Each blue dot in the graph represents an individual property. It is positioned according to its square meterage and corresponding total price. The collection of these points forms a scatterplot shape, indicating a trend of data concentration.

The red regression line is visually dominant and it is a statistical tool. It indicates the best linear fit to the data points based on the least squares method. This line indicates the average effect of the size of a property on its price, demonstrating the direct relationship that the larger the size of a property, the higher the price.

A gray shaded area around the regression line indicates the standard error of the estimates. This shaded band highlights the variability and uncertainty inherent in the regression estimates. It reflects the extent to which actual prices deviate from the predicted prices based on square meters alone.

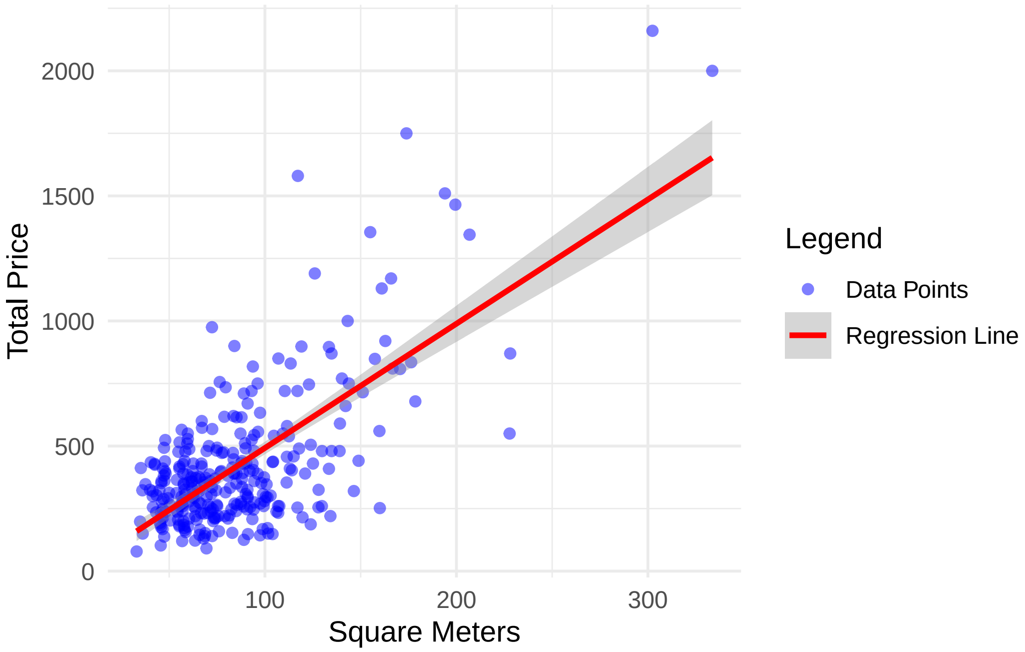


Figure 6: Linear Regression of Total Price on Square Meters

4.3 Model summary

```
# A tibble: 4 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    32.3      68.2       0.473 6.36e- 1
2 square          4.90      0.312    15.7   7.62e-41
3 floor           1.44      2.32      0.619 5.37e- 1
4 buildingType  -17.1      14.0      -1.22  2.23e- 1
```

The linear regression model offers a quantitative analysis of how various factors such as square footage, floor level, and building type impact the total price of houses in Beijing. For instance, the coefficient for square footage indicates a positive relationship, suggesting that as the size of the property increases, so does the price, which aligns with common real estate market expectations. Similarly, higher floors might command higher prices due to preferred views and reduced noise, reflected in the positive coefficient for the floor variable. The impact of building type, quantified through its coefficient, reveals how different types of constructions compare in terms of market valuation, potentially showing preferences for modern over traditional styles, or high-rises over bungalows, depending on the coefficient values.

Additionally, the R-squared value of the model, which measures the proportion of variance in house prices explained by the independent variables, provides an overall assessment of the model's explanatory power. A higher R-squared value would indicate that the model does an excellent job in capturing the variations in house prices based on the included features. This metric, along with the significance levels (p-values) of the individual coefficients, helps in evaluating the reliability and robustness of the model. If certain variables show non-significant p-values, they may not be as influential in determining house prices as initially thought, prompting further investigation or reconsideration of these factors in the model. Thus, the linear regression analysis not only aids in understanding specific factor influences but also in assessing the overall effectiveness of the model in explaining real estate pricing dynamics in Beijing.

4.4 New dataset

After all these analysis, I create a new dataset considering the price change in the house market of Beijing Appendix A.2.

5 Discussion

5.1 Things we have done in this paper

In this paper, I provide a comprehensive analysis of Beijing's housing market, focusing on the relationship between the size of a home in square meters and its total selling price. My goal is to understand the drivers of property values in one of the most populous cities in the world. I use a dataset containing several years of sales data to analyze the impact of property attributes on pricing.

In this paper, we meticulously clean the dataset by removing all non-numeric characters and irrelevant data points to ensure the integrity of the analysis. I then graph the data to visualize trends and patterns to pave the way for more formal econometric modeling.

In this paper, we construct and refine a linear regression model to make it a predictive tool for analyzing the Beijing housing market. First, I identify hypothetical variables that affect housing prices based on economic theory and previous empirical studies. The main independent variable I focus on is the size of the property in square meters because it is a recognized determinant of housing value. In addition to area, other control variables are included to account for factors such as property floor and building type, categorical variables that reflect the diversity of urban housing options.

5.2 Something we learn about the world

Probably the first important takeaway from this paper is the confirmation of a predictable trend in the real estate market - the larger the area, the higher the price of the property. This finding coincides with conventional wisdom, but grounding it in statistical analysis adds weight to the empirical analysis. It shows how market dynamics respond predictably to property characteristics, providing insights into how consumers value space in urban settings.

Second, as the difference near the regression line shows, area alone cannot explain price changes. This suggests that there are other factors at play in determining property values, potentially leading to a discussion of the desirability of location, market volatility, or other attributes unrelated to square footage that influence buyers' decisions.

5.3 Weaknesses

The analysis in this paper, while reliable in many respects, has its limitations. One obvious weakness is the model's reliance on historical sales data, which may not fully reflect current or future market trends. The dynamic nature of the real market means that past relationships between variables do not necessarily apply to the future, especially in the context of rapid urban growth and changing economic conditions.

Another limitation is the assumption of linearity inherent in regression models. Real-world data often exhibit nonlinear relationships, and while linear regression is a powerful predictive tool, it may oversimplify the complex interactions between the various factors that determine house prices. In addition, our model may suffer from omitted variable bias because other influences such as the quality of nearby schools, the economic status of the neighborhood, or even intangible factors such as architectural style or neighborhood ambiance may not be included in the dataset.

The model also assumes that the relationship between square meters and total price is constant across all size and price levels, which may not be the case in a heterogeneous market like Beijing. For example, the relationship between price per square meter may be different for luxury homes versus ordinary homes.

In addition, this paper does not focus on spatial analysis, which is particularly relevant to the real market because geographic location can have a huge impact on property values. Spatial autocorrelation, where the value of a property is correlated with the value of neighboring properties, was not addressed in this paper, which could bias the results and lead to biased coefficient estimates.

5.4 Next steps

The insights gained from the current analysis of the Beijing housing market still have much to be expanded. One area for further research is the integration of additional predictor variables to capture the multifaceted nature of real estate valuations, such as neighborhood crime rates, green space proximity and accessibility to public services. Incorporating these factors could lead to a better understanding of differences in housing prices.

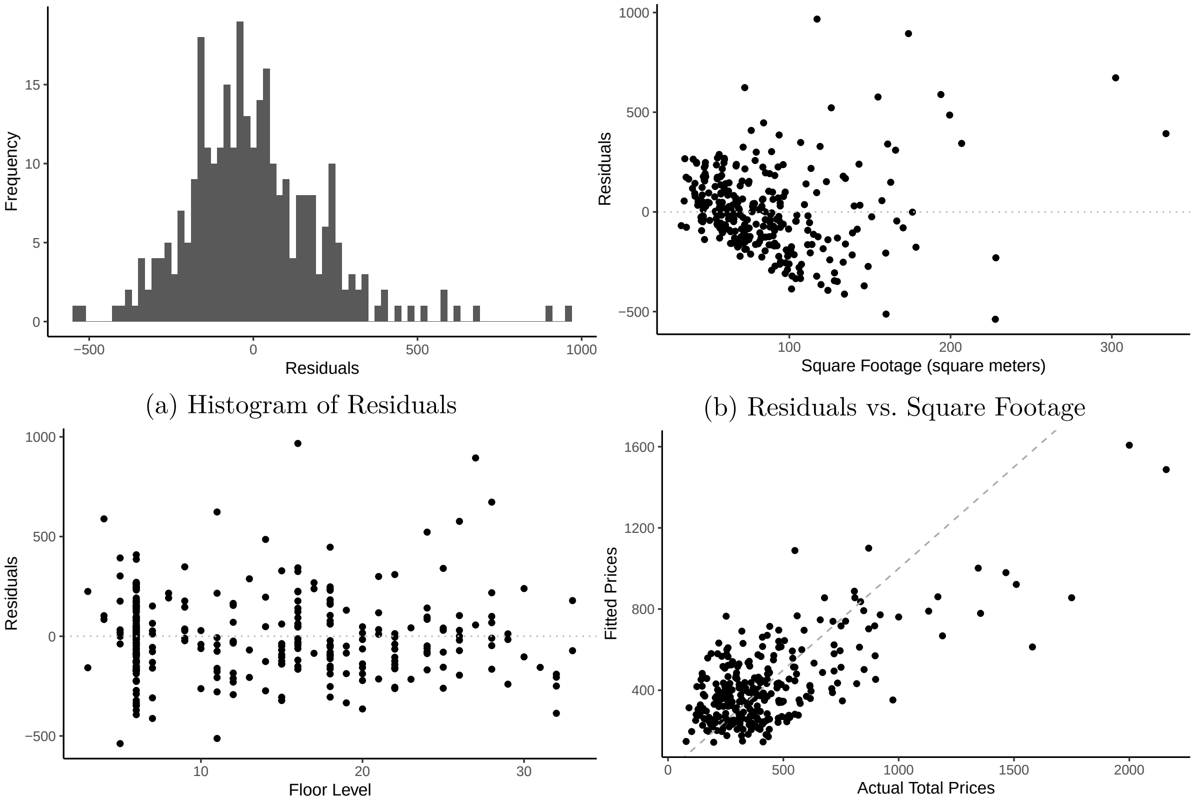
Advances in geospatial analysis offer another frontier for exploration. Given the possibility of spatial autocorrelation, the application of geographically weighted regression models can assess how location-specific factors affect housing prices. These models can reveal how the desirability of certain areas drives the market and how this influence changes across the urban landscape.

Incorporating temporal dynamics is also critical. Real estate markets change in response to economic cycles, urban development policies, and demographics. Valuable predictive insights can be gained through longitudinal studies that examine housing prices over time, particularly in response to policy changes or major economic events.

Finally, there is room to explore the behavioral aspects of real estate transactions. Qualitative research on the motivations, value perceptions, and decision-making processes of buyers and sellers can complement quantitative models and provide a comprehensive view of the market. This human dimension is particularly important in a rapidly urbanizing context such as Beijing, where housing is not only an economic commodity but also an important component of social welfare.

A Appendix

A.1 Model details



(a) Histogram of Residuals

(b) Residuals vs. Square Footage

(c) Residuals vs. Floor Level

(d) Fitted vs. Actual Prices

Figure 7: Residuals from the simple linear regression

In Figure 7: - picture (a): the residuals are normally distributed, centered around zero, indicating a good fit. - picture (b) & (c): These plots help to check for any patterns in residuals across the values of square footage and floor level, respectively. The absence of patterns (e.g., no funnel shape or increasing/decreasing trends) suggests that the model fits well across different values of these predictors. - Picture (d): This plot checks how well the predicted values from the model match the actual values.

A.2 new dataset

The newly created CSV file, new_house_price_dataset.csv, is a curated dataset derived from a larger collection of data concerning the Beijing housing market. This file specifically includes a selection of key variables that are instrumental for analyzing and predicting house prices in Beijing.

References

Liu, Mei, and Qing-Ping Ma. 2021. "Determinants of House Prices in China: A Panel-Corrected Regression Approach." *The Annals of Regional Science* 67 (1): 47–72.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

———. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David. 2014. "Broom: An r Package for Converting Statistical Analysis Objects into Tidy Data Frames." *arXiv Preprint arXiv:1412.3565*.

Wickham, Hadley, R Francois, L Henry, and K Müller. 2014. "Dplyr." *A Grammar of Data Manipulation 2020 [Last Accessed on 2020 Aug 12] Available from*, Rproject.

Wickham, Hadley, Jim Hester, Romain Francois, Jennifer Bryan, Shelby Bearrows, J Jylänki, and M Jørgensen. 2024. "Package 'Readr'." *Read Rectangular Text Data. Available Online: Https://Cran. R-Project. Org/Web/Packages/Readr/Readr. Pdf (Accessed on 23 August 2023)*.

Wickham, Hadley, and Maintainer Hadley Wickham. 2019. "Package 'Stringr'." *Website: Http://Stringr. Tidyverse. Org, Https://Github. Com/Tidyverse/Stringr*.

Wilkinson, Leland. 2011. "Ggplot2: Elegant Graphics for Data Analysis by WICKHAM, h." Oxford University Press.

Xie, Yihui. 2018. "Knitr: A Comprehensive Tool for Reproducible Research in r." In *Implementing Reproducible Research*, 3–31. Chapman; Hall/CRC.

Yan, Jinhai, Lei Feng, and Helen XH Bao. 2010. "House Price Dynamics: Evidence from Beijing." *Frontiers of Economics in China* 5: 52–68.

Zhang, Lei, and Yimin Yi. 2018. "What Contributes to the Rising House Prices in Beijing? A Decomposition Approach." *Journal of Housing Economics* 41: 72–84.

Zhu, Hao, Thomas Travison, Timothy Tsai, Will Beasley, Yihui Xie, and GuangChuang Yu. 2019. "Package 'kableExtra'."