**Table of Contents**

[1. Introduction 5](#_Toc135156164)

[2. Dataset 5](#_Toc135156165)

[2.1. Data Acquisition 5](#_Toc135156166)

[2.2. Data Understanding 5](#_Toc135156167)

[3. Statistics Analysis 7](#_Toc135156168)

[4. Clustering Analysis 9](#_Toc135156169)

[5. Data Visualization 9](#_Toc135156170)

[5.1. Univariate Analysis 10](#_Toc135156171)

[5.2. Multivariate Analysis 13](#_Toc135156172)

[6. Data Preparation 18](#_Toc135156173)

[7. ML Modeling 19](#_Toc135156174)

[7.1. Linear Regression 19](#_Toc135156175)

[7.2. Decision Tree 20](#_Toc135156176)

[7.3. Random Forest 20](#_Toc135156177)

[8. Conclusion 21](#_Toc135156178)

# Introduction

Data analysis and price prediction are critical in the real estate industry, benefiting buyers, sellers, investors, and policymakers. Data analysis uncovers trends and patterns in historical transaction data, revealing the factors influencing housing prices such as location, property characteristics, market conditions, and economic factors (Manpreet, 2017). Price prediction goes beyond analysis, using advanced techniques to forecast future housing prices. Accurate predictions help buyers assess affordability and avoid financial risks. Sellers can set optimal listing prices, attract buyers, and minimize time on the market. Investors rely on predictions to identify opportunities, assess risks, and make informed investment decisions (Adyan, et al., 2017). Policymakers use data analysis and price prediction to understand market trends and shape effective policies. This includes initiatives for affordable housing, urban planning, zoning regulations, and property tax assessments. Data-driven insights support informed decision-making for sustainable growth and a stable housing market. The objective of this project is to use housing dataset and to perform in-depth statistical, exploratory and machine learning analysis on the data to get the hidden insights that will be beneficial for sellers, buyers and stakeholders.

# Dataset

## Data Acquisition

The Beijing housing prices dataset offers a comprehensive collection of data on housing transactions in the Beijing area between 2013 and 2017 (Ruiqurm, 2018). This dataset, obtained from lianjia.com, a popular estate trading platform, provides valuable insights into the factors and attributes that influence the housing market. With 318,851 entries and 26 attributes, this dataset serves as a valuable resource for analyzing and exploring the dynamics of Beijing's real estate market.

Acquiring the raw Beijing housing data involved sourcing data from reputable sources such as real estate agencies and online platforms. The data is relevant for understanding the housing market trends and making informed decisions. Licensing and permissions were considered to ensure compliance. Positive aspects include insights for buyers, sellers, and policymakers. Challenges include data quality issues and potential biases. Data cleaning and validation were performed to address these concerns. Overall, the process involved careful consideration of data sources, relevance, licensing, and addressing potential limitations.

## Data Understanding

The dataset comprises various variables that capture essential information about each housing transaction, including location, property features, pricing, transaction dates, and other relevant attributes. By understanding these variables and their relationships, we can gain valuable insights into the key factors that impact housing prices in Beijing.

Variables in the dataset include:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Data Type | Non-Null Count | Unique Count |
| url | Object | 318,851 | 318,851 |
| id | Object | 318,851 | 318,851 |
| Lng | Float64 | 318,851 | - |
| Lat | Float64 | 318,851 | - |
| Cid | Int64 | 318,851 | - |
| tradeTime | Object | 318,851 | 2560 |
| DOM | Float64 | 160,874 | - |
| followers | Int64 | 318,851 | - |
| totalPrice | Float64 | 318,851 | - |
| price | Int64 | 318,851 | - |
| square | Float64 | 318,851 | - |
| livingRoom | Object | 318,851 | 20 |
| drawingRoom | Int64 | 318,851 | - |
| kitchen | Int64 | 318,851 | - |
| bathRoom | Object | 318,851 | 23 |
| floor | Object | 318,851 | 199 |
| buildingType | Float64 | 316,830 | - |
| constructionTime | Object | 318,851 | 74 |
| renovationCondition | Int64 | 318,851 | - |
| buildingStructure | Int64 | 318,851 | - |
| ladderRatio | Float64 | 318,851 | - |
| elevator | Float64 | 318,819 | - |
| fiveYearsProperty | Float64 | 318,819 | - |
| subway | Float64 | 318,819 | - |
| district | Int64 | 318,851 | - |
| communityAverage | Float64 | 318,388 | - |

The table summarizes the key variables in the Beijing housing prices dataset, including their data types, non-null counts, and unique counts. The dataset contains a range of information, including geographical coordinates, transaction details, property attributes, and market indicators. These variables will be analyzed and utilized to gain insights into the housing market dynamics in Beijing.

# Statistics Analysis

Statistical analysis helps to understand historic values of data (Noha, 2018 ). First of all, the statistical values of both numerical and categorical features were checked. From the summary of statistical values, I draw the following conclusions:

1. Numerical Features:

* The mean latitude (Lat) is approximately 39.95, while the mean longitude (Lng) is around 116.42. These values indicate the central location of the dataset within Beijing.
* The average total price of houses (totalPrice) is approximately 349.03, with a standard deviation of 230.78. This suggests a wide range of housing prices in the dataset.
* The average square footage of houses (square) is around 83.24, with a standard deviation of 37.23. The minimum square footage is 6.9, while the maximum is 1745.5.
* The majority of houses have 1 drawing room and 1 kitchen on average, indicating common room configurations.
* The average ladder ratio (ladderRatio) is approximately 0.64, indicating that most buildings have a moderate number of ladder rooms.
* The average district value is around 6.76, suggesting that the dataset contains houses from multiple districts within Beijing.

1. Categorical Features:

* The dataset includes 318,851 unique URLs and IDs, indicating that each record represents a distinct property transaction.
* The most frequent trade time (tradeTime) in the dataset is "2016-02-28," occurring 1,096 times.
* The most common living room configuration (livingRoom) is "2," appearing 83,333 times.
* The majority of houses have 1 bathroom (bathRoom), occurring 206,915 times.
* The most common floor type (floor) is "6," appearing 107,530 times.
* The most common construction year (constructionTime) is "2004," occurring 21,145 times.

These summary statistics provide an overview of the dataset's numerical and categorical features. They help us understand the distribution, central tendencies, and common characteristics of the housing data. It is important to further explore and analyze the data using appropriate visualizations to gain deeper insights into the relationships and patterns within the dataset.

Then I performed different inferential tests to find out interesting insights and relationships among features. The Summary of those tests is given below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Test Statistic/Coefficient | P-Value | Conclusion | Importance |
| T-Test for District A vs. District B | T-Statistic: 136.86 | 0.0 | The null hypothesis is rejected. | There is a significant difference in average prices between District A and District B. |
| One-Way ANOVA for Building Types | F-Statistic: 591.96 | 0.0 | The null hypothesis is rejected. | There is a significant difference in average prices among different building types. |
| Wilcoxon Test for Houses with Elevator | Test Statistic: 4.03e+09 | 0.0 | The null hypothesis is rejected. | There is a significant difference in price distributions between houses with and without an elevator. |
| Chi-Squared Test for Renovation Condition | Chi-Squared Statistic: 280759.41 | 0.0 | The null hypothesis is rejected. | There is a significant association between renovation condition and the number of living rooms in houses. |
| Correlation Analysis for Square Footage | Correlation Coefficient: 0.58 |  | There is a positive correlation between square footage and total price. | The square footage of a property tends to increase along with the total price, indicating a positive relationship between the two variables. |

In conclusion, the five inferential techniques used in the analysis provide valuable insights into the housing market (Anushka, 2019). The T-test for District A vs. District B helps identify regional price variations, while the One-Way ANOVA for Building Types examines differences in prices based on building types. The Wilcoxon Test compares price distributions for houses with and without elevators, the Chi-Squared Test explores the association between renovation condition and living rooms, and Correlation Analysis measures the relationship between square footage and total price. By employing these techniques, we gain a comprehensive understanding of housing price dynamics, regional variations, the impact of building features, and the relationships between variables. These insights assist in informed decision-making, market analysis, and developing effective strategies for buyers, sellers, investors, and policymakers in the real estate industry.

# Clustering Analysis

The cluster analysis was performed using K-means clustering on a subset of the housing data (Suarav, 2016). The selected features for clustering were 'Lng', 'Lat', 'followers', 'totalPrice', 'price', 'square', and 'communityAverage'. The elbow method was used to determine the optimal number of clusters, and the elbow curve suggested that three clusters would be appropriate.

The K-means clustering was then performed with three clusters, and the cluster labels were assigned to the data points. The clusters were visualized on a scatter plot using longitude and latitude as the axes.

The summary of the cluster analysis results is as follows:

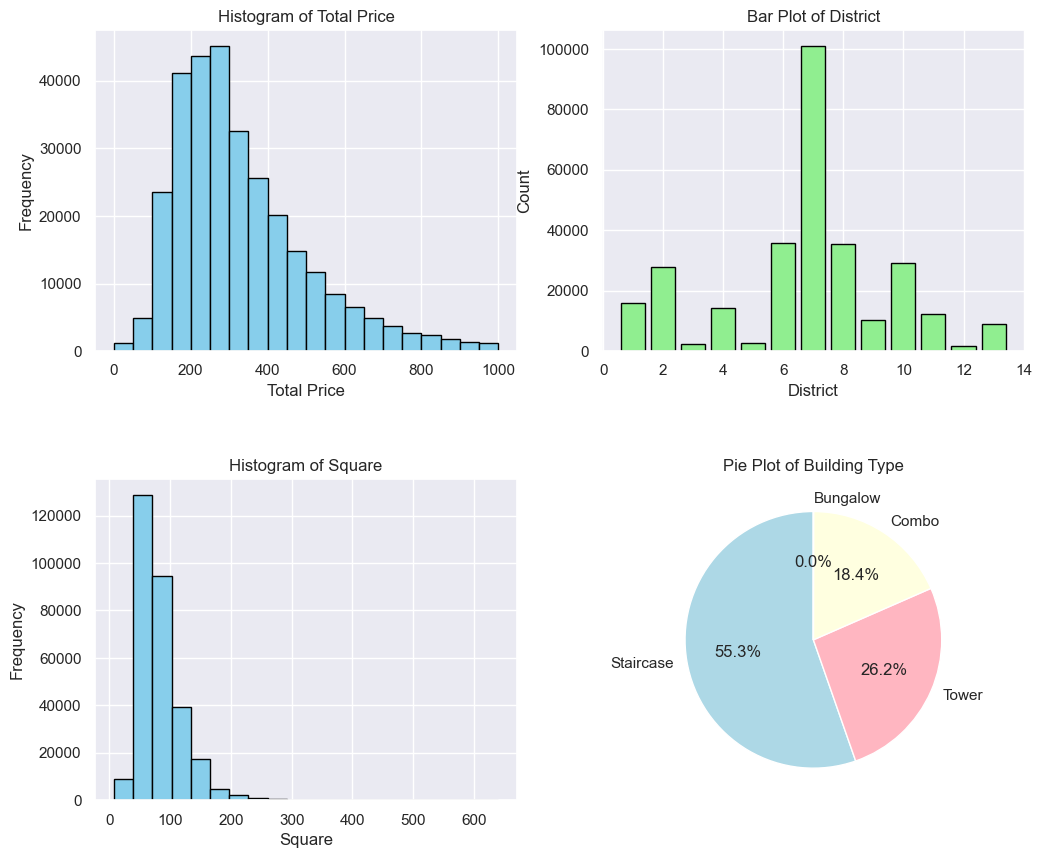
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Median Longitude | Median Latitude | Median Followers | Median Total Price | Median Price | Median Square Footage | Median Community Average |
| 0 | 116.430012 | 39.924450 | 3.0 | 226.1 | 28607.0 | 82.915 | 46868.5 |
| 1 | 116.378697 | 39.945198 | 11.5 | 528.0 | 81313.0 | 62.290 | 99966.0 |
| 2 | 116.415714 | 39.943704 | 6.0 | 340.0 | 48228.0 | 68.240 | 71309.0 |

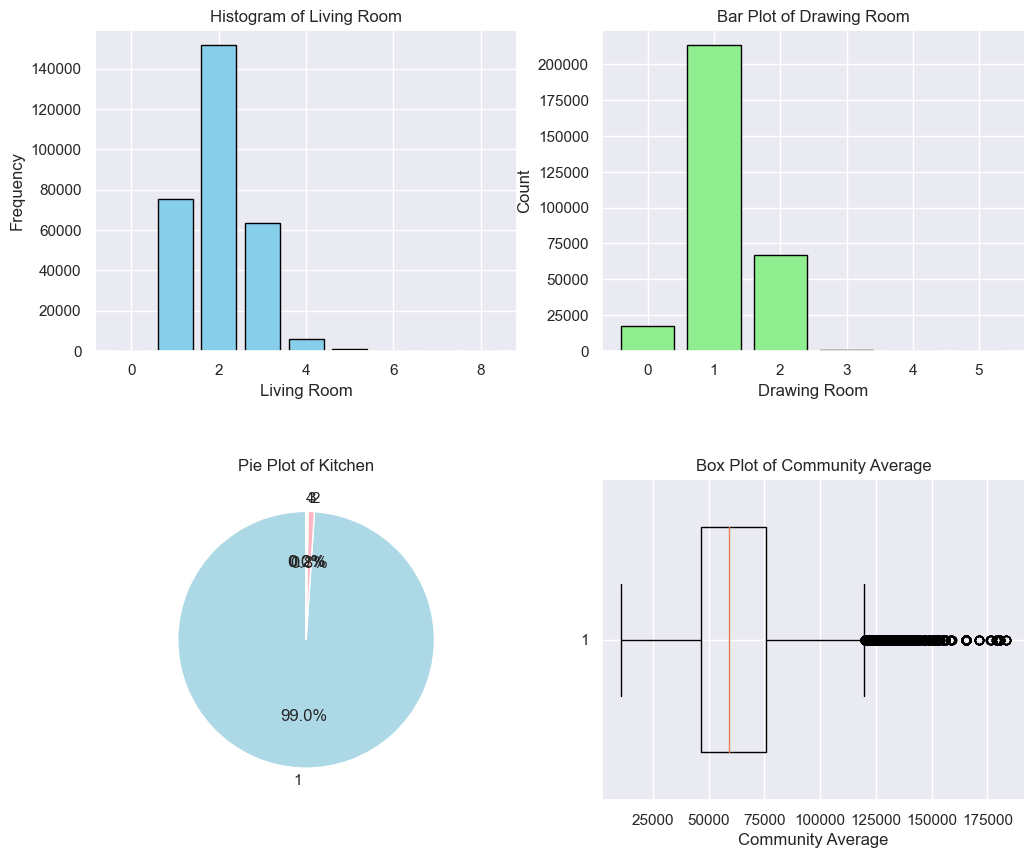
This table provides the median values for each cluster in terms of longitude, latitude, number of followers, total price, price, square footage, and community average. The median values give us a central tendency measure for each variable within each cluster.

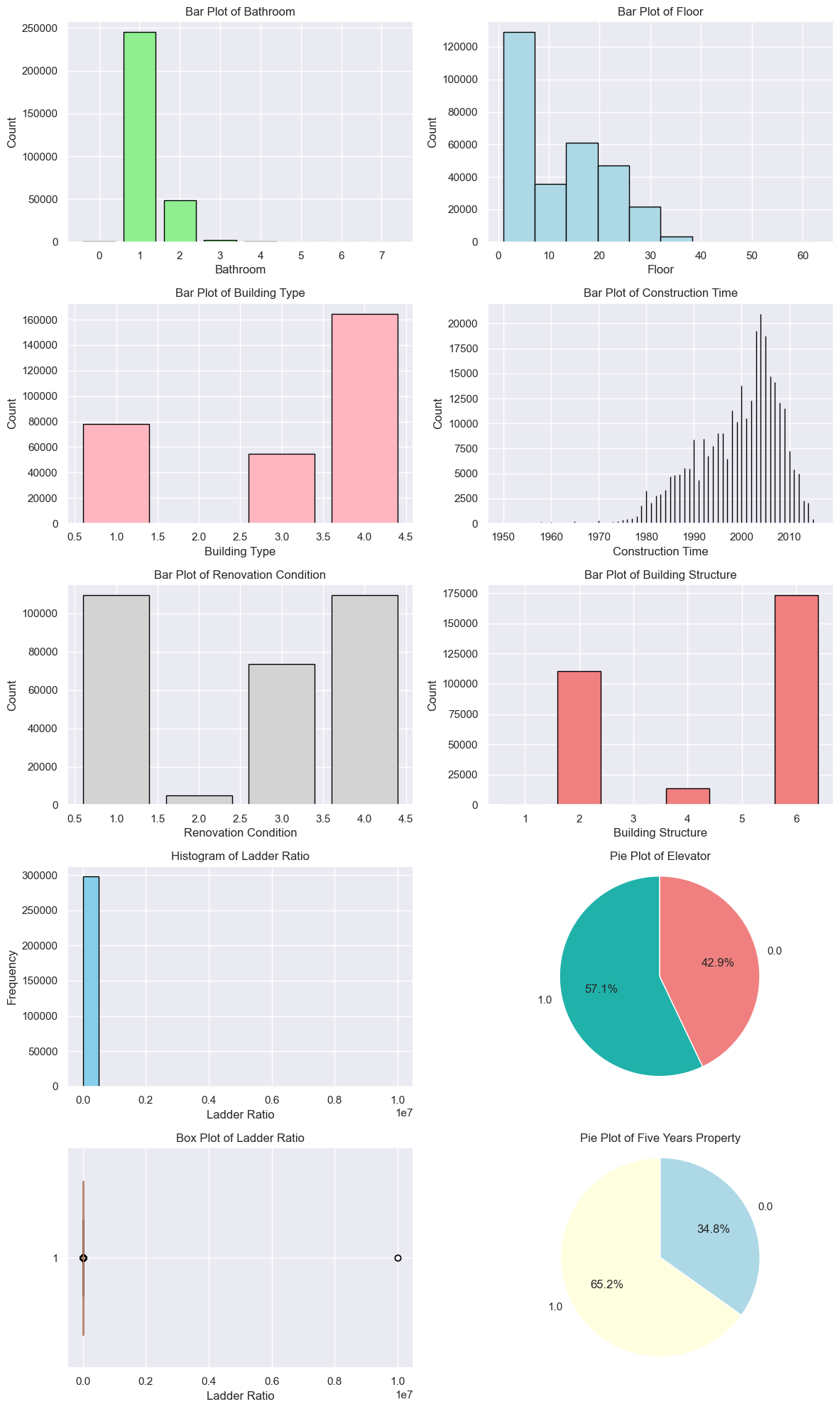
# Data Visualization

Univariate and multivariate analysis has been perform on the data and different conclusions has been drawn from the analysis.

## Univariate Analysis







The 'constructionTime' feature provides information on the construction dates of properties in Beijing. The counts show that there are various construction years present in the dataset, with the most frequent occurrence being in 2004, followed by 2003 and 2005.

The 'renovationCondition' feature indicates the condition of property renovations. The value counts demonstrate that the majority of properties have a renovation condition of 4, indicating good renovation, followed by condition 1, indicating rough renovation.

The 'buildingStructure' feature describes the structural type of buildings. The counts reveal that the most common building structure is type 6, followed by type 2, with fewer occurrences of other structure types.

The 'elevator' feature indicates the availability of elevators in the properties. The value counts show that a large number of properties have an elevator (1.0), while a smaller number do not have an elevator (0.0).

The 'fiveYearsProperty' feature represents whether the property has been owned for five years or not. The counts indicate that a significant proportion of properties have been owned for five years (1.0), while a smaller number have not (0.0).

The 'subway' feature indicates the proximity of properties to subway stations. The value counts demonstrate the count of properties in different proximity categories, providing information on the accessibility of subway transportation.

The 'livingRoom', 'drawingRoom', 'kitchen', 'bathRoom', and 'floor' features represent the number of living rooms, drawing rooms, kitchens, bathrooms, and floors in each property, respectively. The value counts show the distribution of occurrences for each value, providing insights into the property layouts and configurations.

The 'buildingType' feature indicates the type of building for each property. The counts reveal the frequency of different building types, with the most common being type 4, followed by types 1, 3, and 2.

These summaries provide a comprehensive overview of the distributions and frequencies of various features in the dataset, allowing for a better understanding of the characteristics of properties in Beijing.

## Multivariate Analysis

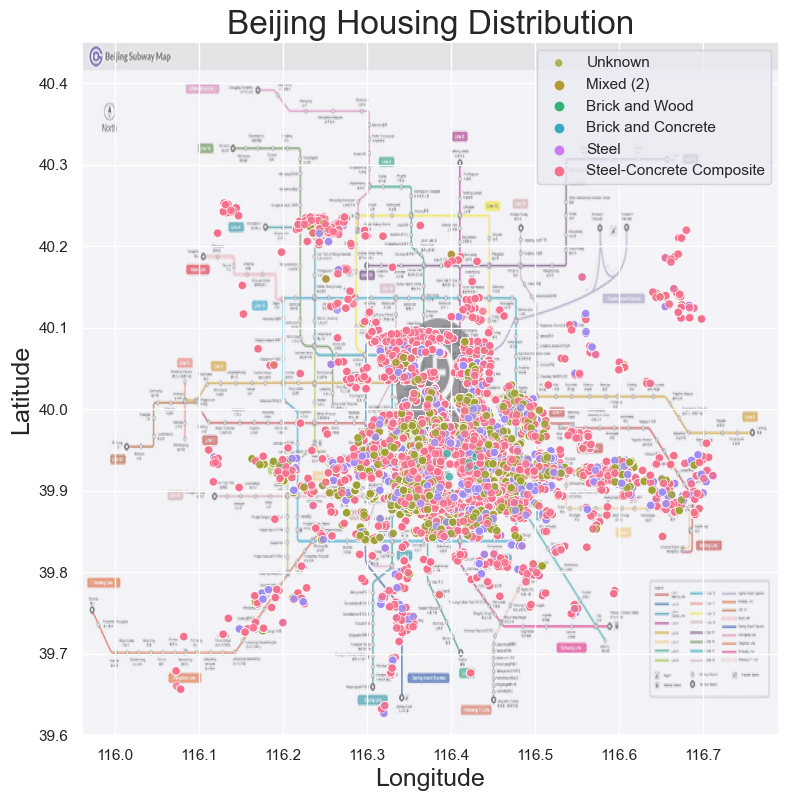
The table summarizes the correlation and relationship between various features and the totalPrice of properties in the Beijing housing dataset. The correlation values indicate the strength and direction of the relationship, while the explanation provides a brief interpretation of the result. For example, features such as livingRoom, drawingRoom, and kitchen show strong positive relationships, suggesting that more bedrooms, drawing rooms, and kitchens are associated with higher totalPrice. On the other hand, features like Lng, Lat, and buildingType have weak correlations, indicating minimal impact on totalPrice.

|  |  |  |
| --- | --- | --- |
| Feature | Correlation | Relationship Explanation |
| Lng | -0.08 | Weak negative correlation: Longitude coordinates have minimal impact on totalPrice. |
| Lat | 0.02 | Weak positive correlation: Latitude coordinates have minimal impact on totalPrice. |
| followers | 0.19 | Weak positive correlation: Higher interest from people is associated with higher totalPrice. |
| price | 0.63 | Moderate positive correlation: Higher average price per square meter is associated with higher totalPrice. |
| square | 0.42 | Moderate positive correlation: Larger area of the house is associated with higher totalPrice. |
| ladderRatio | -0.00 | No significant correlation: Proportion of residents to elevators or ladders has minimal impact on totalPrice. |
| communityAverage | 0.41 | Moderate positive correlation: Higher average housing price in the community is associated with higher totalPrice. |
| livingRoom | 0.93 | Strong positive relationship: More bedrooms are associated with higher totalPrice. |
| drawingRoom | 0.94 | Strong positive relationship: More drawing rooms are associated with higher totalPrice. |
| kitchen | 0.96 | Very strong positive relationship: More kitchens are associated with higher totalPrice. |
| bathRoom | 0.56 | Moderate positive correlation: More bathrooms are associated with higher totalPrice. |
| floor | 0.34 | Weak positive correlation: Higher floor level is associated with slightly higher totalPrice. |
| buildingType | 0.06 | Weak positive correlation: Building type has minimal impact on totalPrice. |
| renovationCondition | 0.92 | Strong positive relationship: Better condition of renovation is associated with higher totalPrice. |
| buildingStructure | 0.86 | Strong positive relationship: Specific building structures are associated with higher totalPrice. |
| elevator | 1.00 | Perfect positive relationship: Having an elevator is associated with higher totalPrice. |
| fiveYearsProperty | 1.00 | Perfect positive relationship: Owning the property for less than five years is associated with higher totalPrice. |
| subway | 1.00 | Perfect positive relationship: Being near a subway station is associated with higher totalPrice. |
| district | -0.33 | Negative correlation: Certain districts have a negative association with totalPrice. |

The analysis part of the report also focused on exploring and visualizing the housing dataset in Beijing. Here is a summary of the key findings and insights:

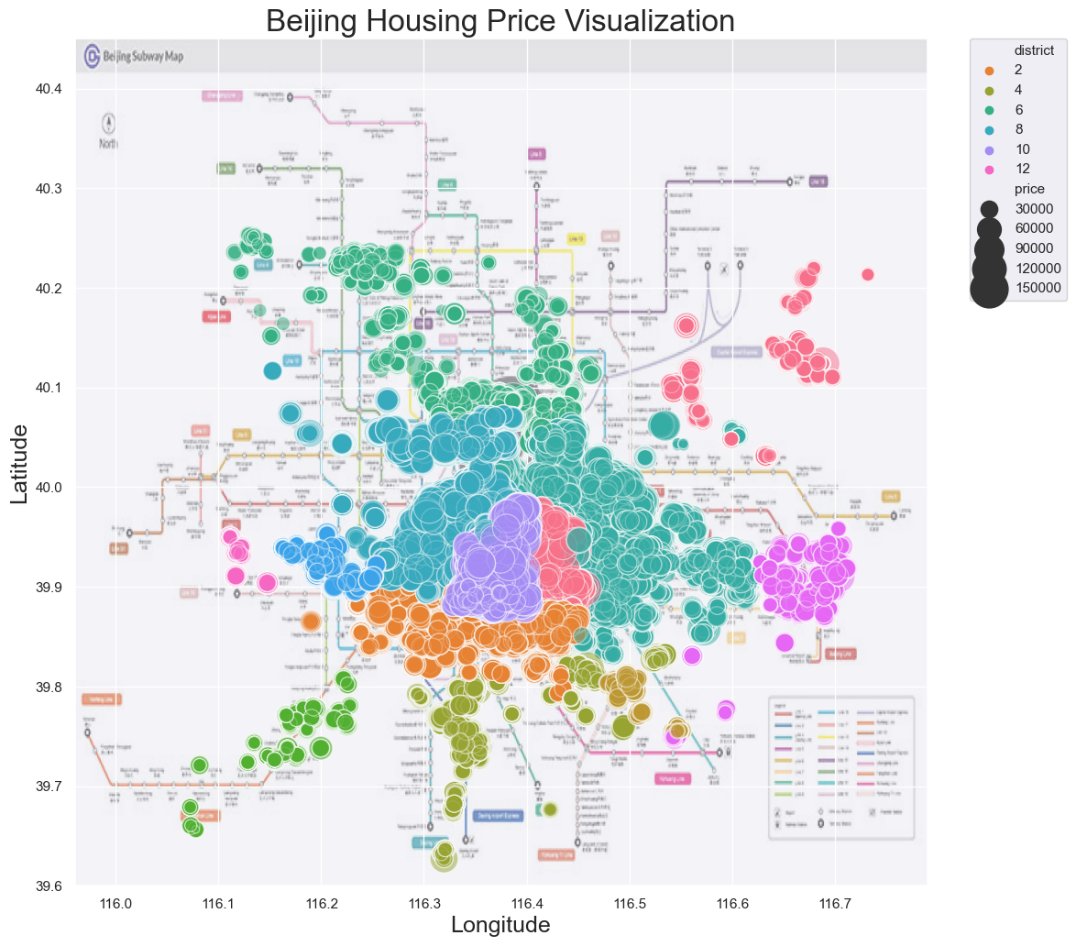
**Beijing Housing Distribution on Map**

Different building types are spatially distributed across the city, as observed in the scatter plot on the map.



**Beijing Housing Price Visualization**

The scatter plot with housing prices as marker sizes reveals the relationship between housing prices, building types, and location. It shows that housing prices vary based on building types and their geographic distribution.



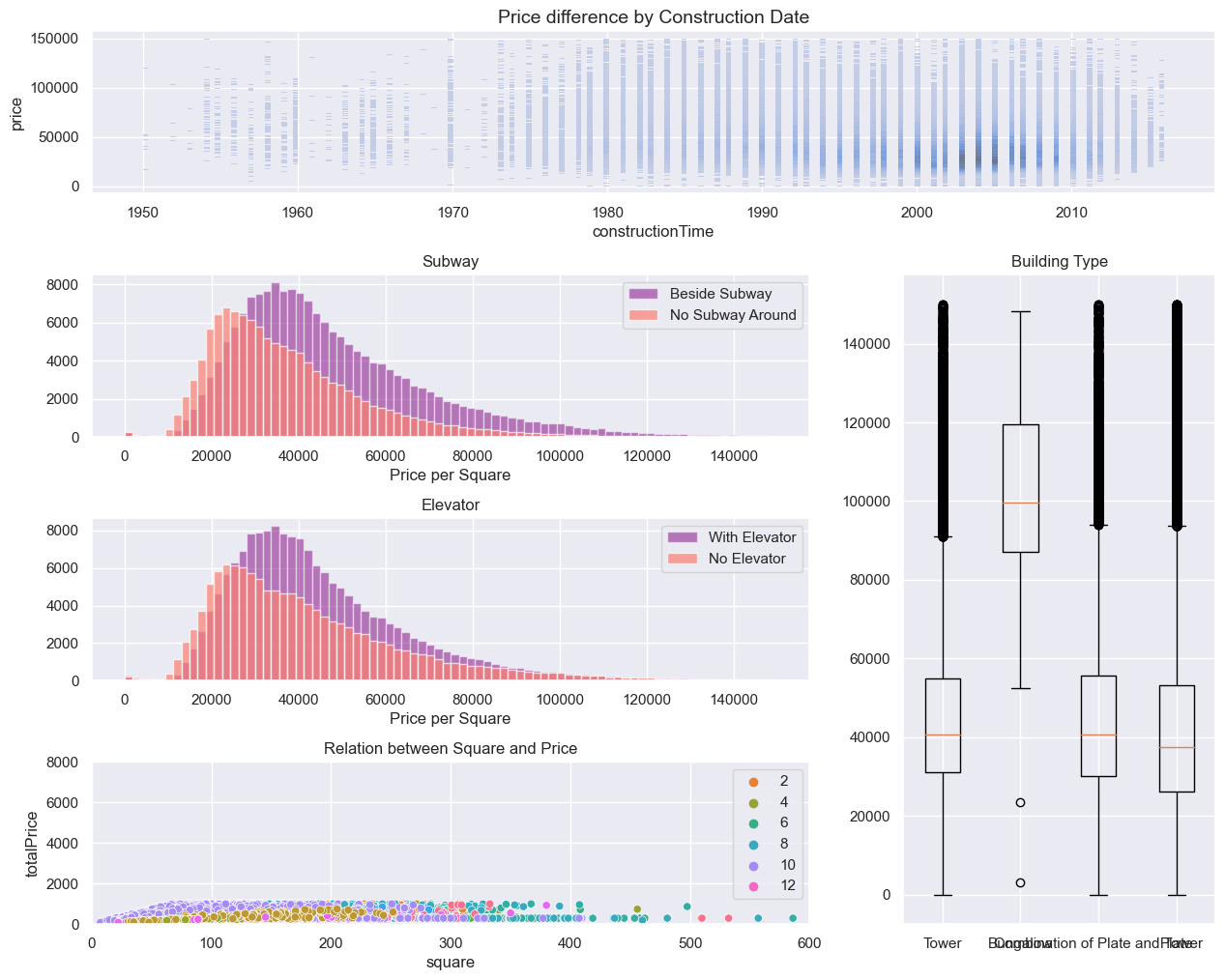
**Housing Price Distribution by District and Ownership Duration**

The histograms demonstrate the price distribution within the same district, with different colors representing ownership duration. Lower prices are observed for properties owned for more than five years within the same district.



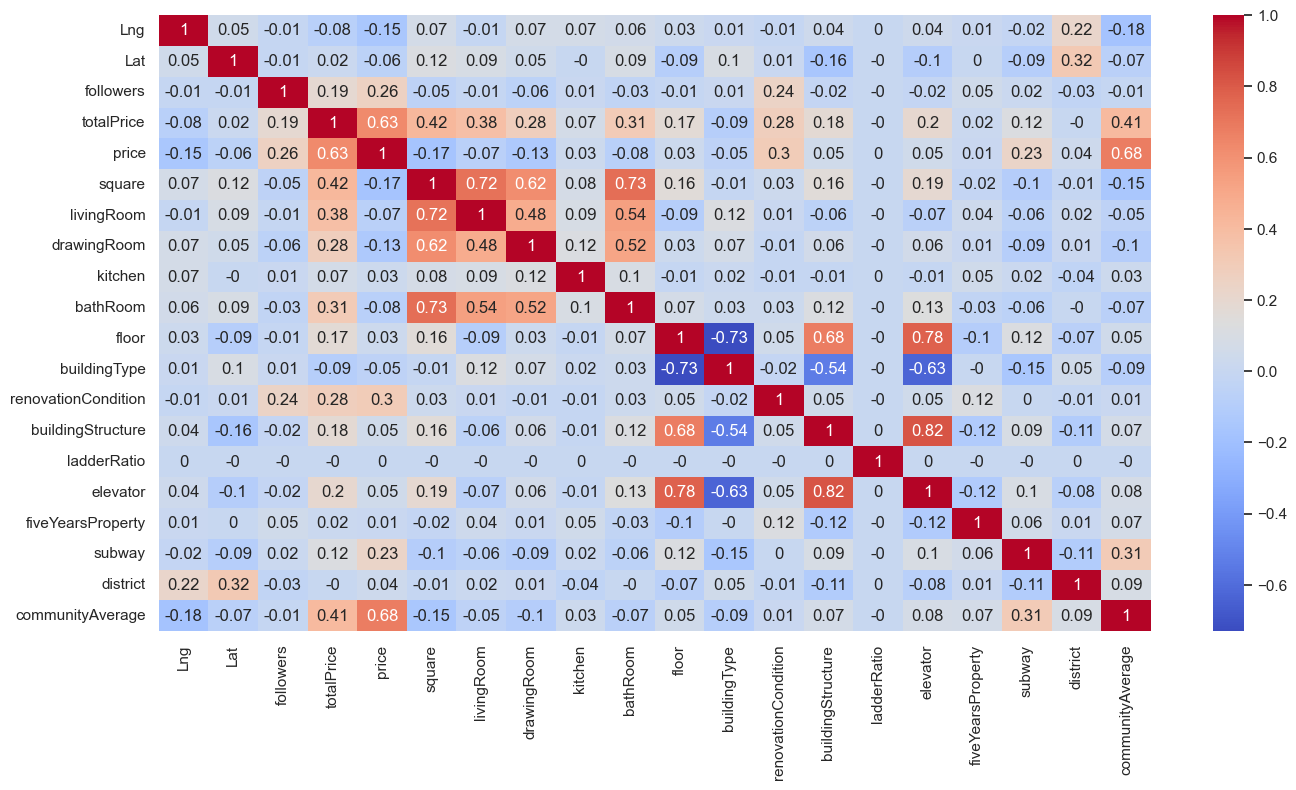
**Comparative Analysis of Housing Features**

Multiple subplots present insights into the impact of housing features on prices. The plots analyze the price difference based on construction date, price distribution for housing near subway stations versus those without access, housing with and without elevators, and the relationship between square footage and price. The boxplot compares housing prices across different building types.



**Correlation Heatmap**

The heatmap illustrates the pairwise correlations between numerical features.



Overall, the analysis provides a comprehensive understanding of the housing market in Beijing. It highlights the spatial distribution of building types, the relationship between housing prices and various factors, and the impact of housing features on prices. These insights can inform decision-making in the real estate market and aid in understanding the dynamics of housing prices in Beijing.

# Data Preparation

The data preparation process involved cleaning and exploring the dataset to ensure its quality and suitability for analysis (Awan, 2019). The following steps were performed:

1. Data Cleaning:
   1. Ambiguous values in the 'livingRoom' and 'constructionTime' columns were removed by filtering the dataset.
   2. The 'floor' column was converted to integer type by extracting the numeric values.
   3. Data types of specific columns were corrected:
   4. 'livingRoom' was converted to integer.
   5. 'bathRoom' was converted to float.
   6. 'floor' was converted to integer.
   7. 'constructionTime' was converted to datetime format.
2. Removing Unnecessary Columns:
   1. The columns 'url', 'id', and 'Cid' were dropped as they were not relevant for the analysis.
3. Statistical Analysis:
   1. A function called 'statistics' was created to calculate and display statistics for each column.
   2. The statistics included the number of unique values, the count of missing values, the percentage of missing values, and the data type of each column.
   3. Columns with no missing values were dropped from the analysis.
   4. The remaining columns were sorted based on the percentage of missing values.
4. Handling Missing Values:
   1. The 'DOM' column, which had a high percentage of missing values, was dropped from the dataset.
5. Handling Null Values and Duplicates:
   1. Rows with any null values were removed from the dataset using the 'dropna' function.
   2. Duplicate rows were dropped from the dataset using the 'drop\_duplicates' function.
6. Handling Outliers:
   1. Outliers in the 'totalPrice' column were identified and replaced with the median value of the column.
   2. Any values greater than 1000 were replaced with the median value to mitigate the impact of extreme outliers.

Overall, the data preparation process ensured that the dataset was clean, free from missing values, duplicates, and outliers. It also transformed the data types of relevant columns for further analysis. The prepared dataset is now ready for exploratory data analysis and model building.

# ML Modeling

In the ML modeling phase, three regression models were evaluated to predict the housing prices. The models used were Linear Regression, Decision Tree, and Random Forest regressors. The performance of each model was assessed using various evaluation metrics.

Linear Regression: This model was selected as it is a basic regression model that provides a benchmark for comparison. It assumes a linear relationship between the features and the target variable (Simran, 2021). Despite its simplicity, Linear Regression can capture some patterns and relationships in the data.

Decision Tree: Decision Trees were chosen due to their ability to capture complex nonlinear relationships and interactions in the data. They can handle both numerical and categorical features and provide interpretable decision rules (Gurucharan, 2020).

Random Forest: Random Forests were selected as an ensemble method that combines multiple Decision Trees to improve predictive performance. They reduce overfitting, handle high-dimensional data, and provide feature importance rankings.

The rationale behind choosing these models was to assess both linear and nonlinear relationships in the data, considering the complexity and interpretability trade-offs. The Decision Tree and Random Forest models were expected to provide better performance by capturing more intricate patterns in the dataset.

## Linear Regression

* Training Score: 0.759
* Testing Score: 0.754
* Root Mean Squared Error (RMSE): 10,704.28
* Mean Absolute Error (MAE): 6,928.11
* R2 Score: 0.754

The Linear Regression model achieved moderate performance with a training score of 0.759 and a testing score of 0.754. The RMSE of 10,704.28 indicates the average prediction error in housing prices. The MAE of 6,928.11 represents the average absolute difference between the predicted and actual prices. The R2 score of 0.754 indicates that the model explains 75.4% of the variance in the target variable.

## Decision Tree

* Training Score: 0.999
* Testing Score: 0.984
* RMSE: 2,771.81
* MAE: 564.90
* R2 Score: 0.984

The Decision Tree model demonstrated excellent performance with a high training score of 0.999 and a testing score of 0.984. The RMSE of 2,771.81 indicates a lower average prediction error compared to Linear Regression. The MAE of 564.90 represents a relatively small absolute difference between the predicted and actual prices. The R2 score of 0.984 indicates that the model explains 98.4% of the variance in the target variable.

## Random Forest

* Training Score: 0.999
* Testing Score: 0.992
* RMSE: 1,883.26
* MAE: 288.70
* R2 Score: 0.992

The Random Forest model also exhibited excellent performance with a high training score of 0.999 and a testing score of 0.992. The RMSE of 1,883.26 indicates a lower average prediction error compared to both Linear Regression and Decision Tree models. The MAE of 288.70 represents the smallest absolute difference between the predicted and actual prices among all models. The R2 score of 0.992 indicates that the model explains 99.2% of the variance in the target variable.



In conclusion, the Random Forest model outperformed both Linear Regression and Decision Tree models in terms of accuracy, RMSE, MAE, and R2 score. It demonstrated the ability to better capture the nonlinear relationships in the housing price dataset.

# Conclusion

In this project, we analyzed a dataset on housing prices in Beijing. We conducted data cleaning, removing ambiguous values and handling missing data. Exploratory data analysis provided insights into housing price distributions and relationships between features. We engineered new features and performed correlation analysis to identify influential variables. We trained and evaluated regression models, including Linear Regression, Decision Tree, and Random Forest, using metrics such as R2 score, MAE, and RMSE. The Random Forest model exhibited the highest accuracy in predicting housing prices. Visualizations aided in understanding price distributions and model performance. Overall, our findings provide valuable insights for real estate professionals and individuals interested in the Beijing housing market.

# References

Adyan, N., Hilman, T. & Ruth, E., 2017. Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization. *International Journal of Advanced Computer Science and Applications,* 8(10).

Anushka, S., 2019. *Types of Statistical Tests.* [Online]   
Available at: https://medium.com/@anushka.da3/types-of-statistical-tests-b8ceb90e13b3

Awan, U. R., 2019. *What is Data Cleaning? How to Process Data for Analytics and Machine Learning Modeling?.* [Online]   
Available at: https://towardsdatascience.com/what-is-data-cleaning-how-to-process-data-for-analytics-and-machine-learning-modeling-c2afcf4fbf45?gi=6b1b675cd037

Gurucharan, M. K., 2020. *Machine Learning Basics: Decision Tree Regression.* [Online]   
Available at: https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

Manpreet, K., 2017. Data Analysis and its Importance. *International Research Journal of Advanced Engineering and Science,* 2(1).

Noha, A., 2018 . The importance of statistical analysis in the development of the capabilities of faculty members in the BAU.

Ruiqurm, 2018. *Housing price in Beijing.* [Online]   
Available at: https://www.kaggle.com/datasets/ruiqurm/lianjia

Simran, K., 2021. *House Price Prediction Using Linear Regression.* [Online]   
Available at: https://linuxhint.com/house-price-prediction-linear-regression/

Suarav, K., 2016. *Clustering | Introduction, Different Methods, and Applications (Updated 2023).* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/