

# Beijing House Pricing Prediction

## 1. Problem Statement

### The Problem

The Beijing housing market is highly competitive. Figuring out the best selling price is the number one decision to make when someone is going to sell a house. The purpose of the project is to build models to predict house pricing in Beijing China.

### Potential Client

Real estate agencies can be the potential clients. A house pricing prediction model with good performance would be very valuable for a real estate agent who could make use of the information provided on a daily basis to decide the house listing price for his or her own clients.

### Source of data

[Kaggle](#) Housing price of Beijing from 2011 to 2017, fetching from Lianjia.com

## 2. Data Wrangling

### Cleaning Steps

- Review column names and rename as needed
- Clean and set index column
- Visualize and find columns have the most missing values
- Convert Chinese characters to English
- For numerical variables:
  - cleaned and performed imputation on missing values
- For categorical variables
  - Split column that has more than one features into separate columns
  - Cleaned and performed imputation on missing values
  - Created labels as needed
  - Converted to categorical dtype where applicable
- Detect and deal with outliers

### Dealing with Missing Values

Missing values mainly appear in four columns: constructionTime, buildingType, communityAverage and DOM (days on market). We will use common sense to decide the best statistics or value to do imputation on missing values.

Because houses that are in the same community tend to be constructed around the same time and with the same building type, we will use the mode for each community to fill the missing values for construction time and building type.

DOM and community average can vary significantly by communities. And the time when the transaction happened would have an impact on those two variables as well. I will use the most recent value by community for each year to impute missing values in community average column, and use the mode by community by year to impute missing values in DOM.

## Detecting Outliers

Outliers with extremely high or low values appear in two columns, ladderRatio and totalPrice.

For ladderRatio column which has extreme high maximum value that seems incorrect, I simply removed the outliers because there are just a couple of lines showing as outliers

For totalPrice column where the values can be validated based on the product of price per sq. meter and total sq. meters of the house, I compared the original totalPrice values with the calculated values. Then removed lines that have significant differences between the two values. After cleaning the data and removing the outliers, we have 316222 lines that are ready to use for the next section. 2622 (0.8%) lines are dropped during the data cleaning process.

Summary of Fields in the Data Set

| Numeric Variables  | Categoricals   | Objects | DateTime  | Targets                       |
|--|--|---------|-----------|-------------------------------|
| Lng<br>Lat<br>Cid<br>DOM<br>followers<br>squareMeters<br>bedRoom<br>livingRoom<br>kitchen<br>bathRoom<br>buildingType<br>constructionTime<br>buildingStructure<br>ladderRatio<br>communityAverage<br>floorPosition<br>buildingFloors | buildingType<br>renovationCondition<br>buildingStructure<br>elevator<br>fiveYearsProperty<br>subway<br>district<br>floorPosition | url     | tradeTime | totalPrice<br>pricePerSqMeter |

### 3. Exploratory Data Analysis & Data Storytelling

After data cleaning and processing, I then performed EDA to visualize distribution of the data and correlation between variables. Also I conducted statistical tests to see if the categorical features (elevator, fiveYearsProperty and subway) have a significant impact on the house pricing.

#### Findings from the EDA

1. Between 2010 to 2017, Beijing house pricing had an increased trend from 2010 to 2016 in all districts. Starting 2017, there was a decrease in pricing
2. House pricing varies in different Beijing districts, where Xicheng and Dongcheng and Haidian districts have the highest price over time.

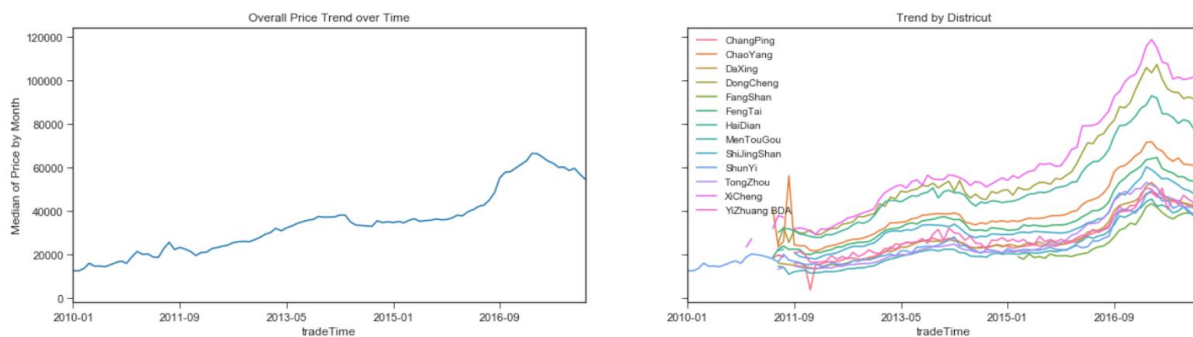


Fig 1 Median of Price Per Square Meter by month (2010-2017)

3. We can see that houses in the urban area or around the center of the city have higher price. Xicheng, Dongcheng and Haidian districts have the most expensive house price medians

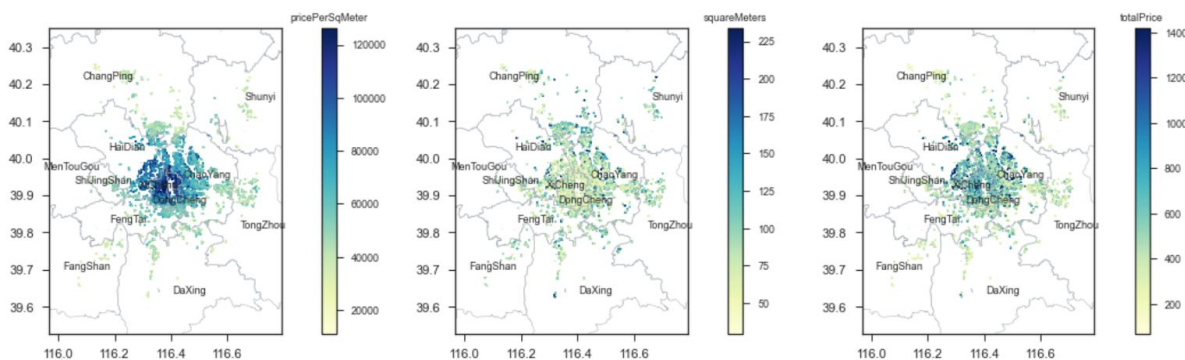


Fig 2 PricePerSqMeter, SquareMeters, totalPrice on Map

4. In terms of building types, bungalow houses have higher price than the other types.
  - a. Bungalow houses typically located in the city center in DongCheng and XiCheng districts where the Tiananmen Square and Forbidden city are located, and have only 1 floor with a brick-and-wood structure. Bungalow houses usually have an older construction time than the other types.
  - b. Tower buildings are the highest building type, mostly over 20 floors. They usually have steel-concrete composite structure

- c. Plate is the most common house building type in Beijing. Mostly have less than 10 floor with mixed building structure
- d. Combination building type mostly has 10-20 floors with steel-concrete composite building structure

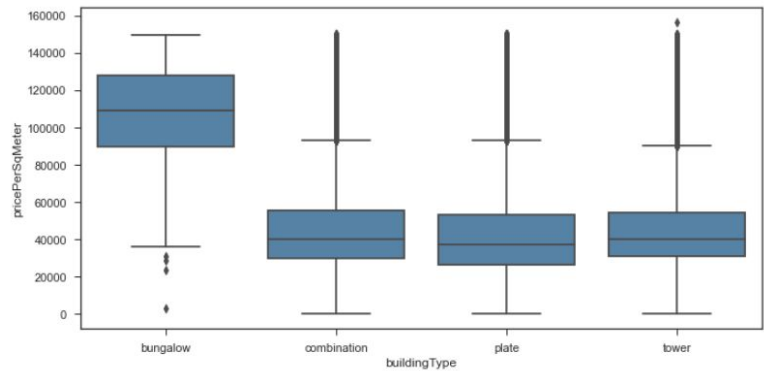


Fig3 PricePerSqMeter by building types

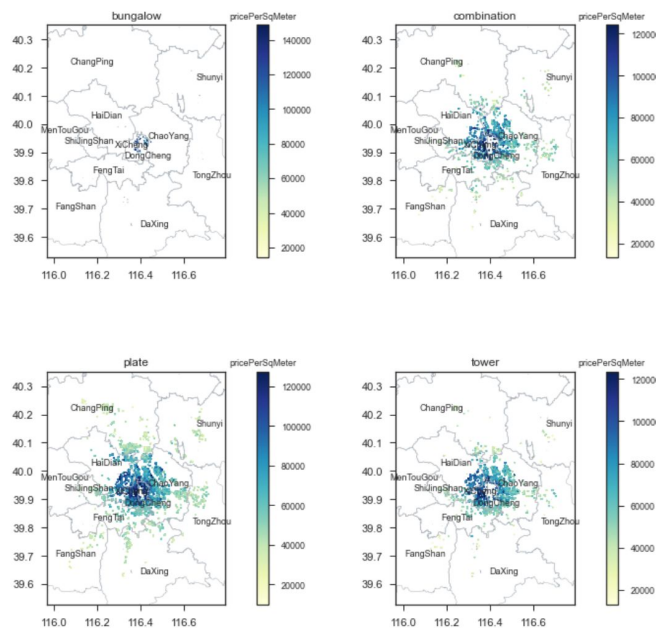


Fig4 PricePerSqMeter by building types on map

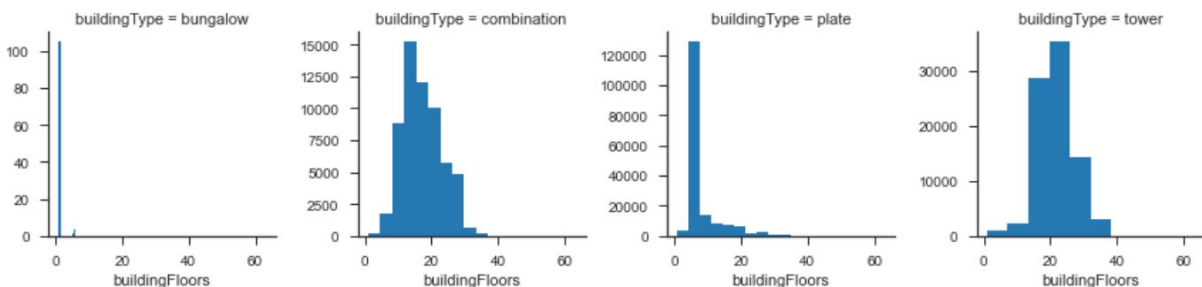


Fig5 Histograms - buildingFloors by building types

5. In terms of renovation condition, houses with "rough" renovation conditions have a lower price than the "basic" and "upgraded" condition which is expected as the "rough"

condition is the lowest level of renovation condition in China. But it seems there is not much difference in pricing for "basic" and "upgraded" renovation conditions.

6. Houses with access to subways have higher price than houses that are away from the subways.

## Statistical Test Summary

| Variable          | Summary of Statistical Tests   |
|-------------------|--|
| elevator          | <ul style="list-style-type: none"><li>• Permutation Test<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Bootstrap Test:<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Mann-Whitney Test<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Welch's T-test<ul style="list-style-type: none"><li>◦ P-value:1.971996620377344e-172</li></ul></li></ul>                   |
| fiveYearsProperty | <ul style="list-style-type: none"><li>• Permutation Test<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Bootstrap Test:<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Mann-Whitney Test<ul style="list-style-type: none"><li>◦ P-value: 2.1382780878680677e-116</li></ul></li><li>• Welch's T-test<ul style="list-style-type: none"><li>◦ P-value: 1.5354933157064842e-14</li></ul></li></ul> |
| subway            | <ul style="list-style-type: none"><li>• Permutation Test<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Bootstrap Test:<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Mann-Whitney Test<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li><li>• Welch's T-test<ul style="list-style-type: none"><li>◦ P-value: 0.0000</li></ul></li></ul>                                  |