Code

Stat 315, Final Report: Beijing Traded Houses Between 2002 and 2018

Kaili Chen

kailic

Due Friday, June 28, 2019 at 11:59 pm

Code

Introduction

Housing price one of the most commonplace topics for people who live in populous cities (see link attached below). I happened to find this dataset from Kaggle, on the traded house in Beijing. With an interest of exploring factors affecting housing prices and learning regional differences of housing types and building structures across districts, I decided to use this dataset for my final project.

(https://www.worldatlas.com/articles/the-10-largest-cities-in-the-world.html
(https://www.worldatlas.com/articles/the-10-largest-cities-in-the-world.html))

Code

Data

Code

Table continues below

Lng	Lat	subway	district	totalPrice	price	square	tradeTime
116.5	40.02	Yes	Chaoyang	4150	31680	131	2016-08-09
116.5	39.88	No	Chaoyang	5750	43436	132.4	2016-07-28

Lng	Lat	subway	district	totalPrice	price	square	tradeTime
116.6	39.88	No	Chaoyang	10300	52021	198	2016-12-11
116.4	40.08	No	Changping	2975	22202	134	2016-09-30
116.4	39.89	Yes	Dongcheng	3920	48396	81	2016-08-28
116.5	39.99	No	Chaoyang	2756	52000	53	2016-07-22

Table continues below

constructionTime	buildingStructure	buildingType	elevator
2005	steel-concrete composite	tower	Yes
2004	steel-concrete composite	tower	Yes
2005	steel-concrete composite	plate	Yes
2008	steel-concrete composite	tower	Yes
1960	mixed	plate	No
2005	steel-concrete composite	plate	Yes

renovationCondition	location
Simplicity	Middle
Hardcover	Middle
Simplicity	Middle
Other	Suburb
Rough	Central District
Simplicity	Middle

The dataset originally has 318851 observations and 26 variables. After selecting several variables I am interested in and removing rows with unknown characters and rows containing NA values, the dataset I am analyzing in this project is reduced to 297990 observations with 14 variables:

Column Name	Description
Lng	longitude coordinates
Lat	latitude coordinates
totalPrice	the total price(¥)
price	the price of the house (¥/m^2)
square	the area of house (m^2)
subway	located close to the subway or not

Column Name	Description
renovationCondition	hardcover, simplicity, rough, and other
buildingStructure	brick and wood, brick and concrete, steel, steel-concrete composite, mixed, unknown
buildingType	tower, plate&tower, plate, and bungalow
constructionTime	when was this house constructed
tradeTime	when was this house being sold
district	the district in which this listing is located
location	central district, middle, or suburb

Goals

- Write a 3-5 sentence paragraph that explains your overall analysis goal is (what are you trying to explore? which relationships do you think exist?).

My goal for this project is to analyze how different features or variables related to houses are related with the housing price. Moreover, I am also interested in looking at the geographical distribution of building types and building structures in the city and how they relate to each other. I think there is a relationship between subway and housing price (unit price and total price), between elevator and housing price (unit price and total price). I also think there is a relationship between location and building types, between location and building structures and between building types and builidng structures.

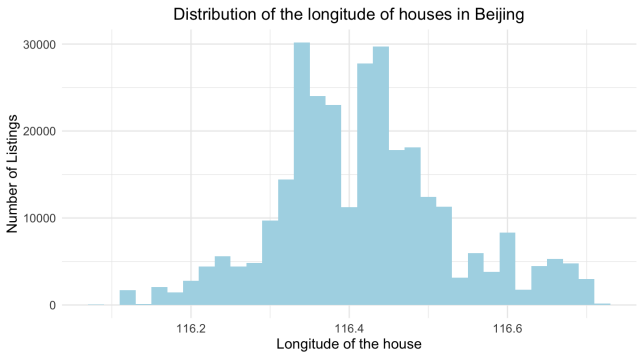
Exploring the variables

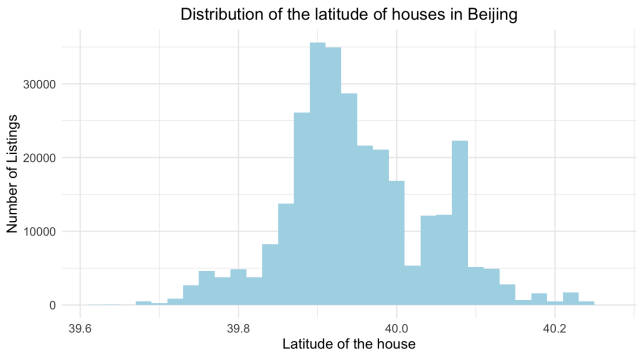
Now, **for each variable**:

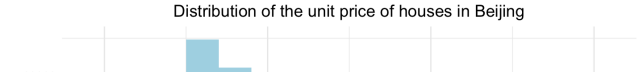
- Create a graph that summarizes its marginal distribution. Feel free to display multiple graphs at a time, keeping in mind the overall aesthetics of your report.
- Describe the graphs (1-2 sentences for each variable).

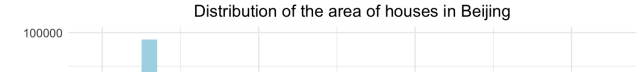
Note: You may want to include some simple, bivariate graphs. This could show any clear relationships that you'll explore further below or show that, for example, many of your variables are highly correlated.

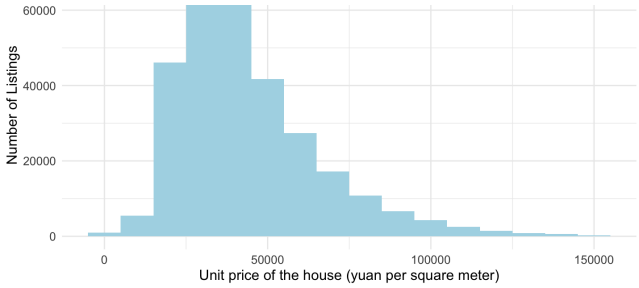
Code

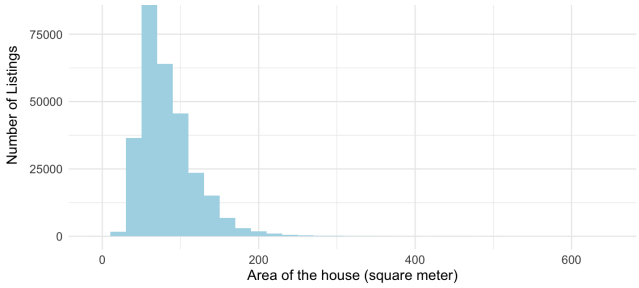


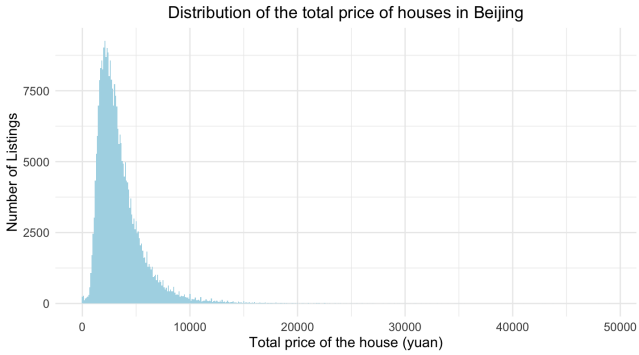


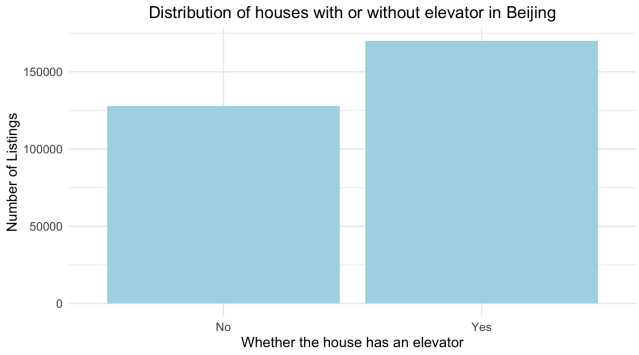


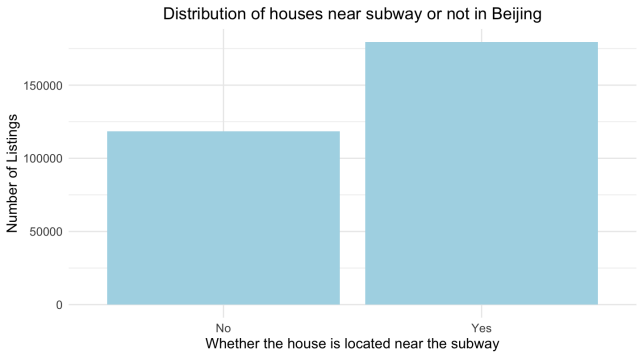


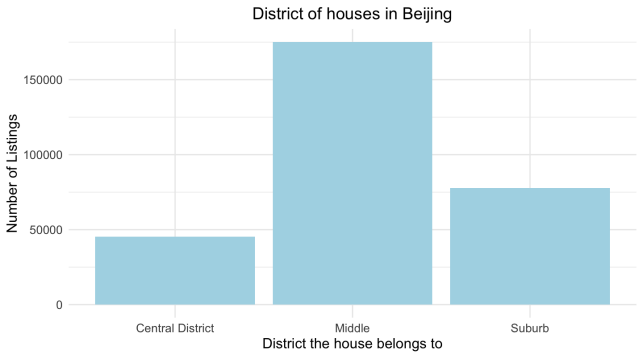


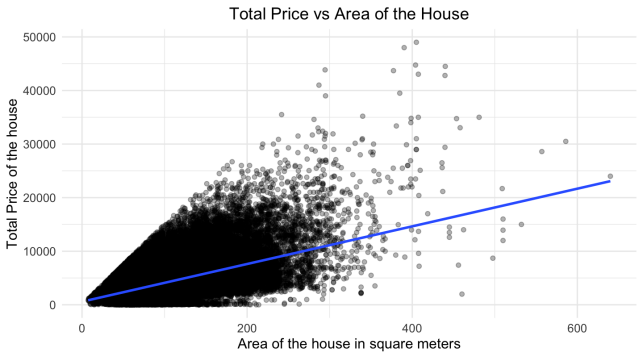


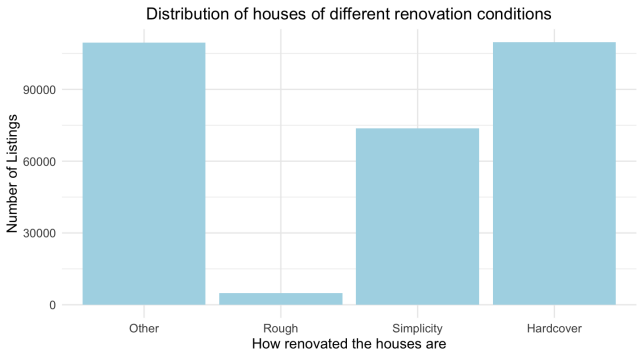


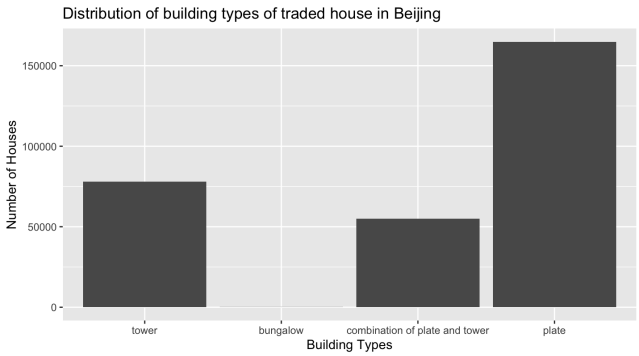


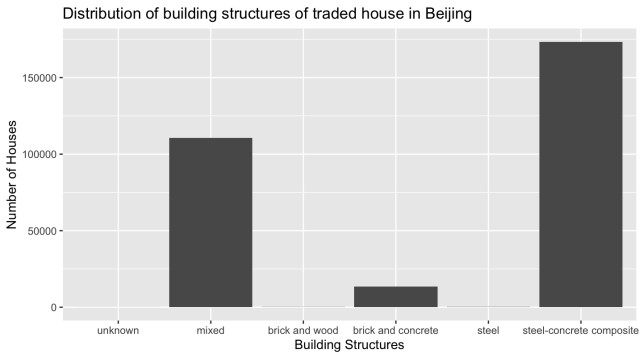












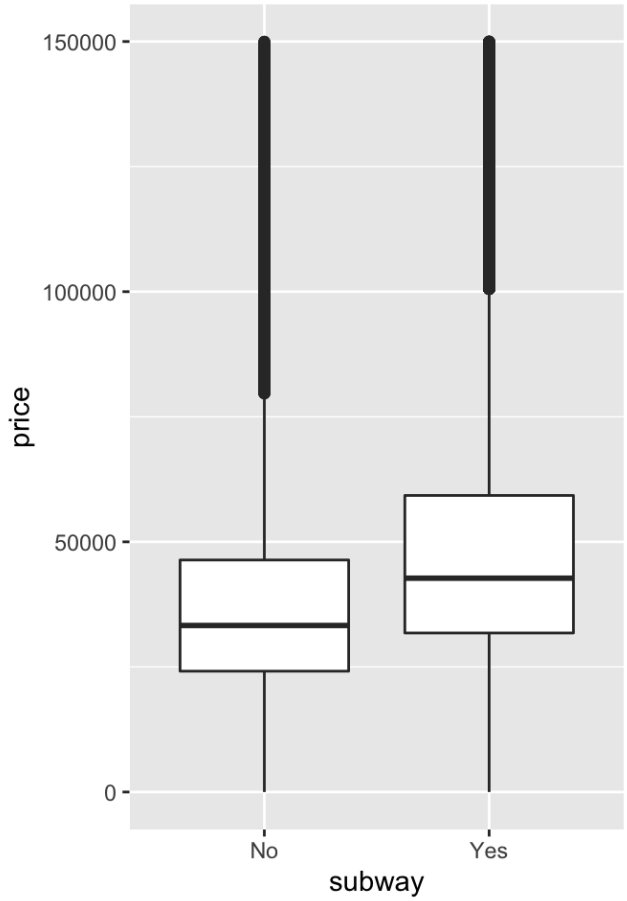
The longitude ranges from 116.1 to 116.7, and the latitude ranges from 39.63 to 40.25. Both longitude and latitude seem to have bimodal shapes. Unit price of houses in Beijing has a minimum of 1 rmb per square meter, a maximum of 150000 rmb per square meter and a mean of 43830 rmb per square meter. The distribution is heavily skewed to the right. Area of houses in Beijing ranges from 7.37 and 640 square meters. Average area of houses is 82.67 square meters. The distribution is heavily skewed

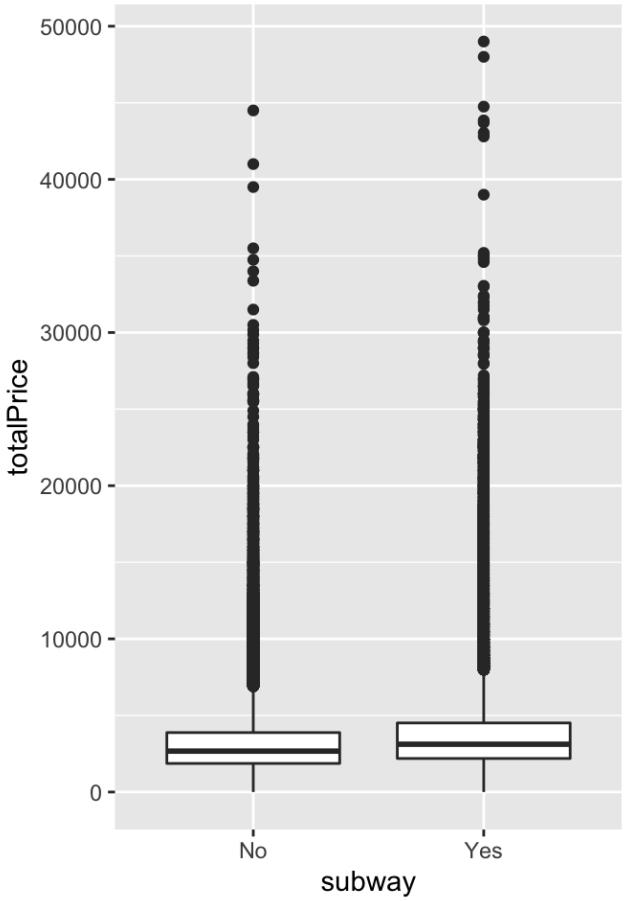
to the right. Total price of houses in Beijing has a minimum of 1000 to 4900k rmb, a mean of 3492k rmb. The distribution is heavily skewed to the right. 127989 houses do not have an elevator in the building and 170001 houses have an elevator. So there are more houses with an elevator than houses without an elevator. 118581 houses are not located close to the subway and 179409 houses are located close to the subway. So more houses are located near the subway than houses located not near the subway. There are 45278 houses in the central district in Beijing, 175030 houses are located in between the central district and suburb of Beijing, 77682 houses are located in suburb of Beijing. The trade time of houses in the dataset I am looking at ranges from 2002-06-01 to 2018-01-28. From the scatterplot, we can see that the area of the house is highly correlated with the total price of the house, which makes sense because larger houses in general are worth of a higher total price.
When looking at the renovation condition of houses, we can see that 4901 houses are roughly renovated and barely furnished, 73784 houses have simple renovation condition and 109683 houses are completely renovated. Time when houses were sold ranges from 2002 to 2018. Its distribution is bimodal with two modes, one between 2012 and 2014, another one between 2015 and 2017. There is a valley around 2014, which means not a lot of houses were sold during that time compared to other times. Building structures of traded houses in Beijing are under four categories: tower, bungalow, combination of tower and plate, and plate. The most common building type in Beijing traded house is plate, and then the next common building type in Beijing traded house is tower and then combination of tower and plate. The least common buildig type is bungalow. Building structures of traded houses in Beijing are under five categories: unknown, mixed, brick and wood, brick and concrete, steel, and steel-concrete composite. Steel-concrete composite is the most common building structure. Mixed is the second most common building structure, followed by brick and concrete. Brick and wood, steel, and unknown building structures are very uncommon.

Methods

Hypothesis 1: Houses that are close to the subway were traded on average at a higher price than houses that are not close to the subway.

Code





Code

```
## [1] 10276.92
```

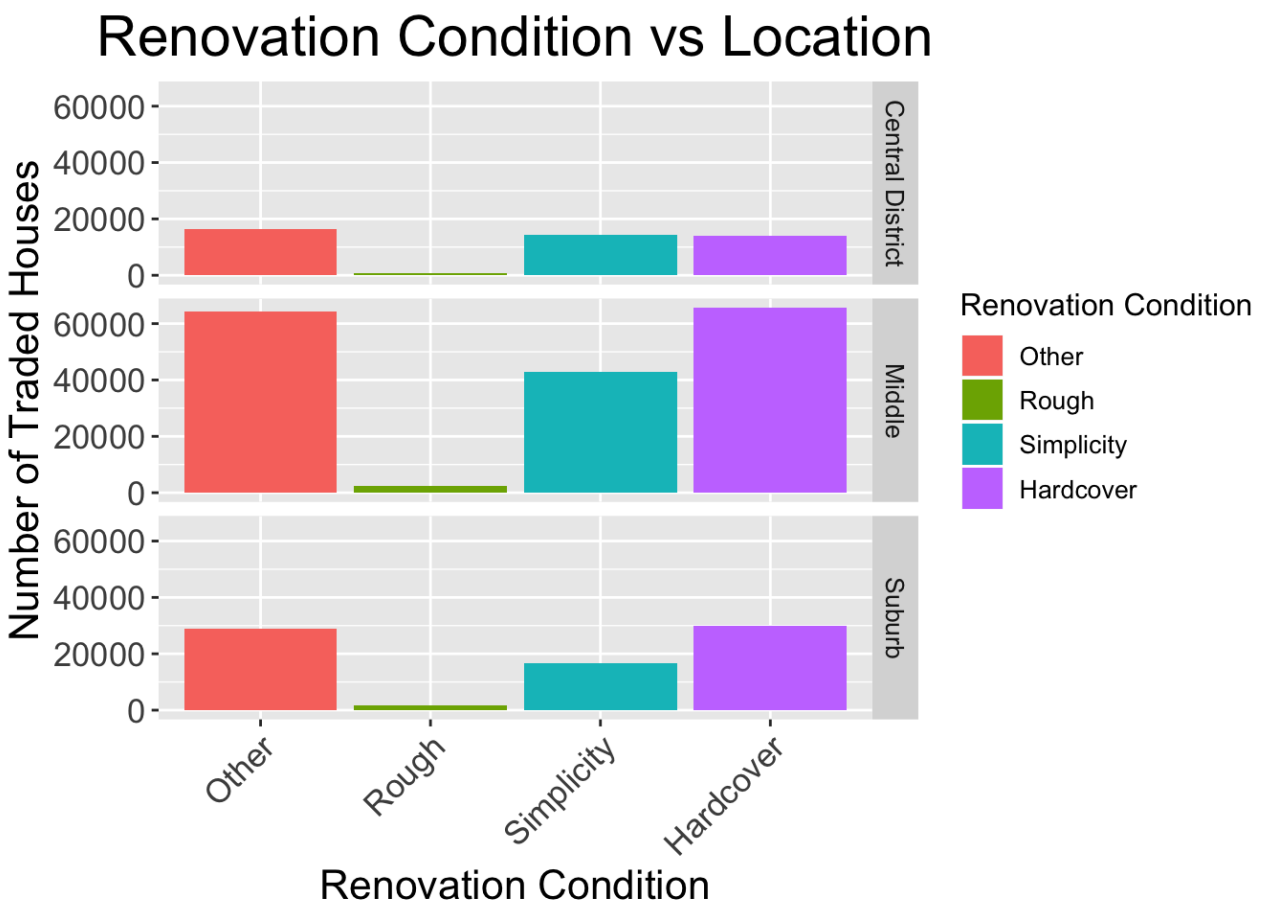
Code

```
## [1] 535.9719
```

To test if my hypothesis is true, I made first made a side-by-side boxplot to compare the unit price of houses located near the subway with houses located far from the subway (the graph on the left). We can tell that the mean of unit price for houses located near the subway is higher than houses located far away from the subway, and they differ by 10276.92 RMB/m^2. Then I made a similar side-by-side boxplot but this time I wanted to compare the total price of houses located near the subway with that of houses located far from the subway. We can tell that the mean of total price for houses located near the subway is higher than houses located far away from the subway, and they differ by 535971.9 RMB.

Hypothesis 2: I think houses in the same district have similar renovation conditions inside of the house.

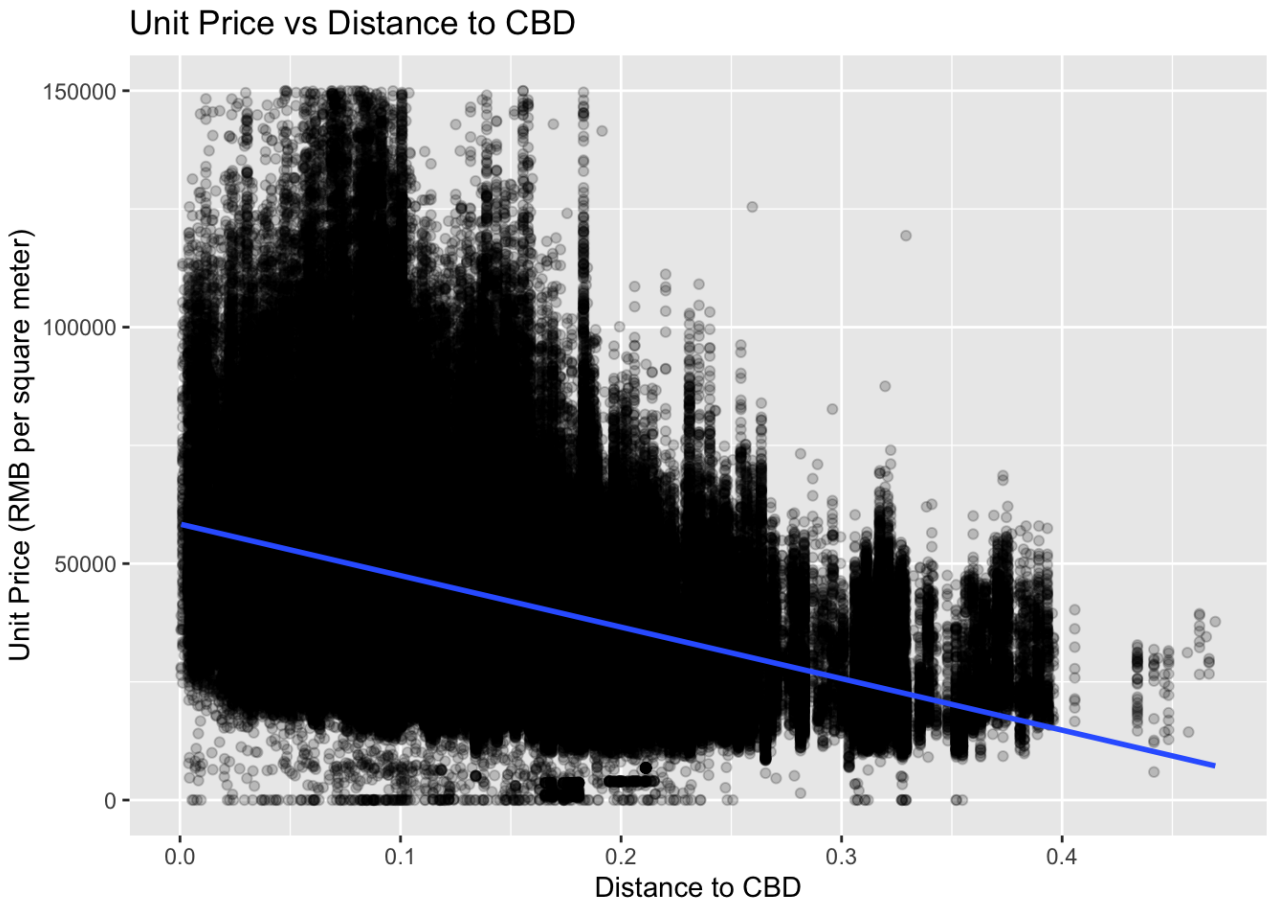
Code



I created three bar charts on the distribution of houses under different renovation conditions, and then I facet them by district to see if there is any difference in different district. My understanding of hardcover, simplicity and rough are houses that are highly renovated and completely furnished, houses that are simply renovated and houses that are barely furnished. As we can see, in central district, there are about similar amount of houses that are highly renovated and houses that are simply renovated. Whereas in the middle area and suburb of Beijing, there are more houses that are highly renovated and completely furnished than houses that are simply renovated. One common thing across all districts is that there are very few houses that are barely furnished or renovated. It makes sense because in such a fast-paced city, there will be a very small market for people who look for houses that are not furnished before moving in.

Hypothesis 3: I think on average unit price increases as the location gets closer to the Central Business District (CBD) of Beijing.

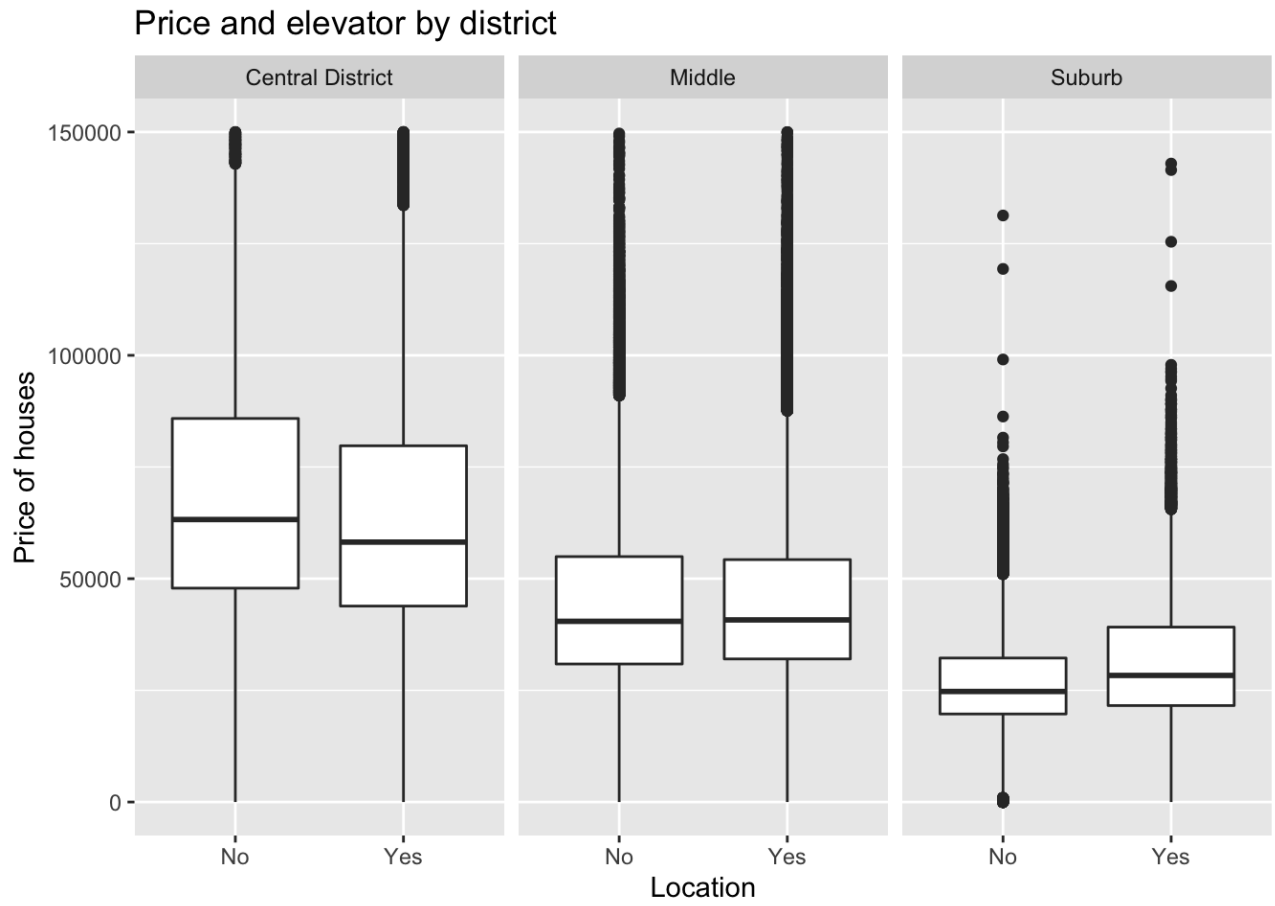
Code



To measure how far each house is away from the center of Beijing, I compute their distances with the coordinates of a point in the central business district of Beijing and saved the distances in a column called `distance_cbd`. After that, I created a scatterplot on price vs distance to CBD (Central Business District) and I added a best fit line to the scatterplot. Since the best-fit linear line has a downward slope, I believe we have enough evidence to conclude that on average, as the distance of the house to CBD increases, housing price on average would decrease.

Hypothesis 4: I think houses with an elevator are associated with a higher unit price for houses across all district in Beijing.

Code



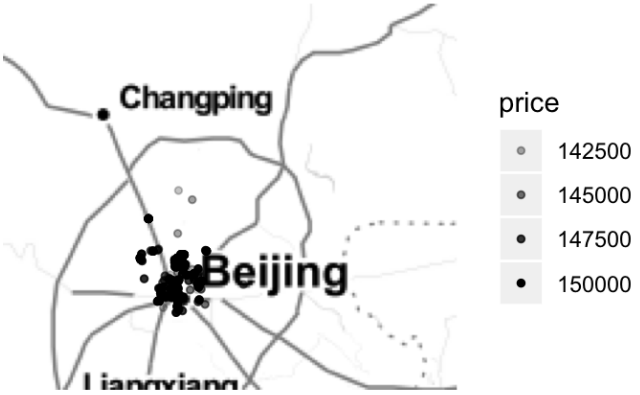
I created three side-by-side boxplots comparing the unit price of houses with an elevator in the building and houses without an elevator, and then I facet the houses by district, to see if there are any regional differences. The result is quite interesting. In central business district, houses with an elevator on average have a lower unit price than houses without an elevator. In the district in between central district and suburb of Beijing, houses with an elevator on average have a similar unit price than houses without an elevator. On the contrary, for houses in the suburb of Beijing, houses with an elevator on average have a higher unit price than houses without an elevator, as I assumed in my hypothesis.

Hypothesis 5:

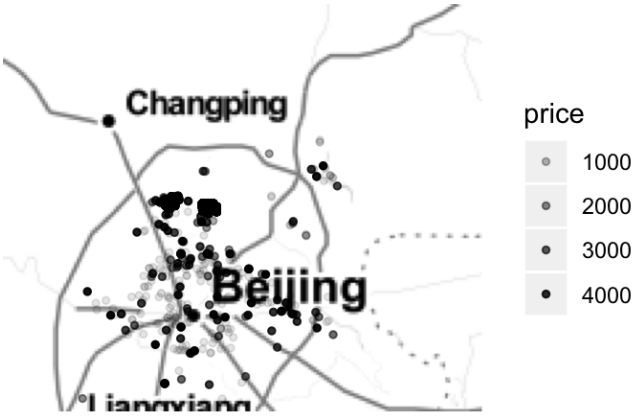
Houses with a relatively high unit price (greater than 140000 RMB per square meter) are located together, so does houses with a relatively low unit price (less than 5000 RMB per square meter).

Code

Houses traded at a relatively high unit price



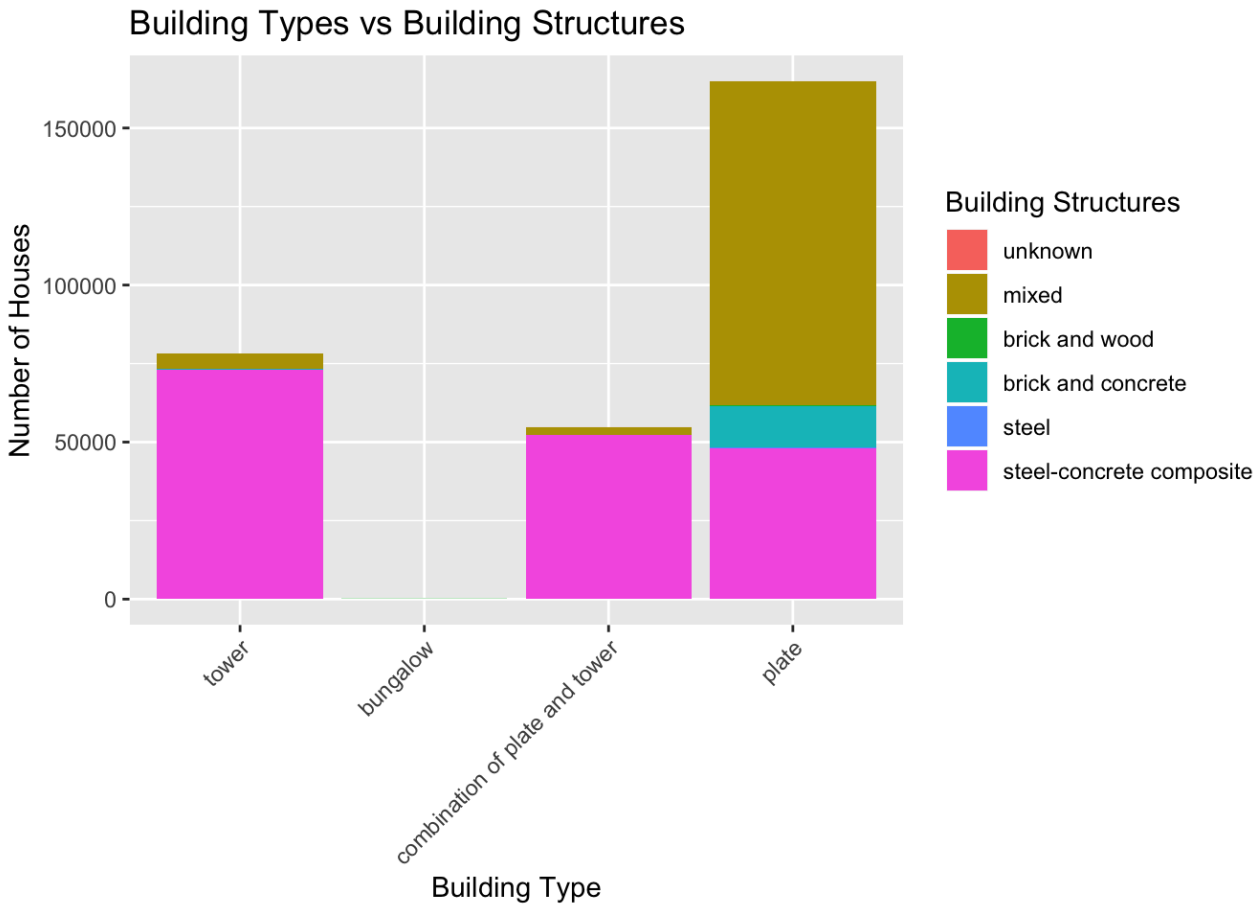
Houses traded at a relatively low unit price



I plotted the geographical location of houses with relatively high unit price, with transparency. The darker the color is, the more expensive the unit price of the house is. I also plotted the geographical location of houses with relatively low unit price, with transparency. Same as the previous graph, the darker the color is, the more expensive the unit price of the house is. From the first graph, we can see that houses traded at a relatively high unit price are clustered in the center of Beijing, with a few houses located somewhere in between the CBD and the suburb. On the contrary, houses traded at a relatively low unit price are spread out pretty evenly across all district in Beijing, with a cluster to the north of CBD.

Hypothesis 6: there is some relationship between building structure and building type.

Code



To see whethere there are any relationship between two categorical variables building types and building structures, I plotted a stacked bar chart of houses of different building types, colored by different building structures. Some interesting takeaway from this graph is that most tower buildings and buildings made of the combination of plate and tower are steel-concrete composite. Most plate buildings are mixed in terms of building structure, and the next common building structure for plate buildings is steel-concrete composite, followed by brick and concrete building structure.

Results

Whether the house is located near the subway is indeed a factor that will affect the average total price and average unit price of houses in Beijing. The distribution of renovation conditions of houses in Beijing differs by district. An increase of distance to CBD is associated with an increase in the unit price. Surprisingly, it's not true that houses with an elevator have on average a higher unit price than houses without an elevator across all district in Beijing. The association between whether the house has an elevator in the building and the unit price of the house varies by districts. I also learned that houses with relatively high unit price are located quite close to each other in the center of Beijing, whereas houses with relatively low unit price are pretty spread out across different areas in Beijing. Lastly, there are some relationship between building types and building materials. For example, most tower buildings and buildings with a type of the combination of plate and tower are made of steel-concrete composite. Brick and concrete appears a lot in plate buildings.

Discussion

Geographically, Beijing has this interesting shape of roads – five rings on a map marked by major roads. It would be interesting to explore whether housing price all shows a five-ring pattern, highest in the middle and then keep decreasing as the ring gets bigger.

One of the biggest limiation of my data analysis is that I haven't had the chance to explore all the 26 variables and how different variables relate to each other. For example, there is a column in the originla dataset called "floor", that indicates the floor of the house. I would like to know the relative height of the house with regard to the building, low, middle, or high.