UNIVERSITY OF CALIFORNIA, SANTA BARBARA

Machine Learning of Real Estate Market in Beijing

Director: Prof. Tomoyuki Ichiba

Report by Meiyu Li

Contents

1 Introduction **1**

2 Data **2**

3 Exploratory Data Analysis **3**

4 Method **8**

5 Results **9**

5.1 Price Distribution by Districts 9

5.2 Prediction - Linear Regression 10

5.3 Prediction - Ridge Regression 11

5.4 Prediction - Lasso Regression 12

5.5 Classification - Logistic Regression 13

6 Conclusion **14**

References **16**

1 Introduction

Beijing is the capital of China, and it is also the commercial, political, and diplomatic center of China. With many potential functions, Beijing attracts many people to study, work, and live in. The housing price in Beijing's real estate market is one of the most concerned topic for most people with the dream of living in Beijing. Besides this, the real estate market is highly significant to the people and leaders of Beijing and the broader world for a variety of political, welfare and economic reasons.

In Beijing, when asking house buyers to describe their dream houses, we may not get answers like the height of the basement ceiling or the proximity to the railroad. Instead, price negotiation including purchase cost, maintenance cost, and financial stability after purchase draw the buyers' most concerns about buying a home. Besides these concerns, house buyers in Beijing also consider the access to transportation, schools nearby, working places as important factors. To address the demand of most house buyers and present how the housing attributes will impact the real estate markets, we choose the dataset that describes the price of individual residential property transaction in Beijing from 2016 to 2018. In this paper, we attempt to answer:

- Along with the local attributes (longitude, latitude, etc.), what do the housing price distribution in Beijing China and in California US present to us, what is the difference of price distribution between two places, and what can we conclude about the impact of local attributes?
- What important factors in the dataset will have significant impact on the real estate market price in Beijing, and how these potential factors will influence the housing price?
- Given the significant variables we found in the Exploratory Data Analysis part, what predictions can we make on housing price by using the regression models we constructed, and what conclusions we will have after comparing these models?

2 Data

As we addressed, we are using dataset HousingPrice.csv (318,851 observations) and there are 26 variables involved in assessing the transaction detail for. These 26 variables focus on the quality and quantity of many attributes relating to each housing transaction.

In total, the variables can be divided into two categories: position attributes and house attributes. Variables like longitude or latitude, whether there is a subway nearby, number of schools or companies nearby, districts, GDP are the variables to best clarify the environmental conditions. There are some variables that indicate the house conditions for each transaction, like the number of floors, whether there is an elevators, number of bathrooms, living rooms, and bedrooms.

This dataset is obtained from Kaggle and provided by the author Qichen QiuQiu (2018). The data comes mainly from LianJia, an authoritative and popular website/APP that gives a transaction platform for home buyers and home sellers in China. Some important attributes are attached below:

HousingPrice.csv

- ID: transaction No.
- Longitude: local attributes
- Latitude: local attributes
- Subway: it indicates if there is a subway nearby.
- Followers: the number of people follow the transaction
- Square: the area of the house(m^2)
- Building Type: the house buildings include types of tower, bungalow, combination of plate and tower, and plate.
- Building Structure: the structure of buildings are listed as unknown, mixed, brick and wood, brick and concrete, steel, and steel-concrete composite.
- Renovation Condition: the condition of the housing renovation is divided into other conditions, rough, Simplicity, and hardcover.
- Construction Time: Overall material and finish quality

- Five Years Property: it indicates if the owner have the property for less than 5 years.

To best evaluate how these attributes impact the sale price, we did the missing values check. We notice that 49.54 percent of DOM (active days on market) are missing values. We decide to omit this nonsense column, since we are unable to know the DOM information for nearly half of the observations and it is hard for us to find an appropriate estimator for the missing values. In addition, we notice that there are around 30 out of 318,851 observations for attributes living room, bathroom, bedrooms, fiveYearsProperty. Since there are only 30 of them are missing and it is hard for us to estimate them, so we simply delete these 30 rows. Finally, there is one column is worth noticing, building Type (categorical). There are 0.63 percent of observations for building types are null values. But there is no missing values in the similar attribute Building Structure (categorical) and there are some rows called "unknown". Instead of deleting the rows containing nulls, we decided to label the missing values as "unknown", like the similar attribute building structure.

3 Exploratory Data Analysis

In Figure 1, We can see the real estate market in Los Angeles appears to be flat from 2008 to 2018. There is no big jump or fall. When we look at the smooth line of Beijing real estate Market, the housing price is generally increasing. In 2011, house prices in Beijing fell by 42.3% due to the Chinese Property Bubble, but prices surged again by 39.5% in 2012. Additionally, there is 29.9% increase in 2016 and the price surge most from 2011 to 2015. To drive the economic expansion in 2016, China depended heavily on the surging real estate market and the government stimulus.

In Figure 2, we can see the housing price in the center of Beijing is the highest and price is decreasing from the center to surrounding area. The red circle lines represent the main roads (1st, 2nd,..., 7th ring roads) in Beijing. The area within the 3rd ring road has more significant commercial, business, and political functions, so the estate market is more expensive and precious within the 3rd ring road than that beyond 3rd ring road. From the Table 1, we can see the price difference between 3rd ring road and 4th ring road is very large. but the price range beyond 3rd ring road does not changing too much. However, noting that the price derived

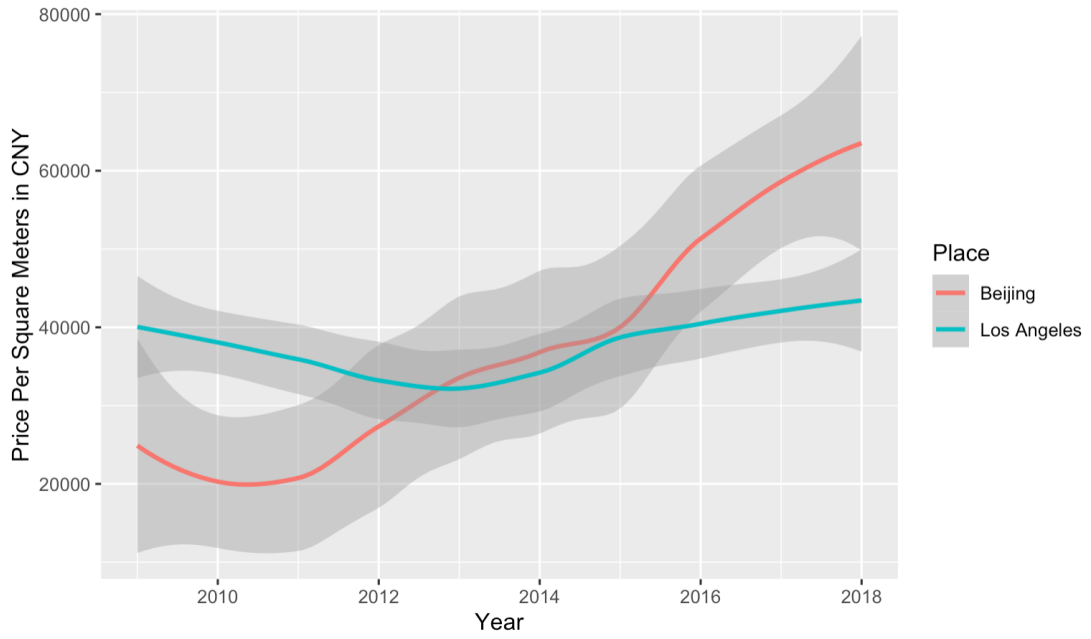

Figure 1: Price(CNY) Per Square Meters in Beijing and CA from 2008 to 2018 (Currency Change Rate Adjusted)

from 2008 to 2018, so the housing price here does not represent current estate market. But we can get a sense of the range between differend rings of roads.

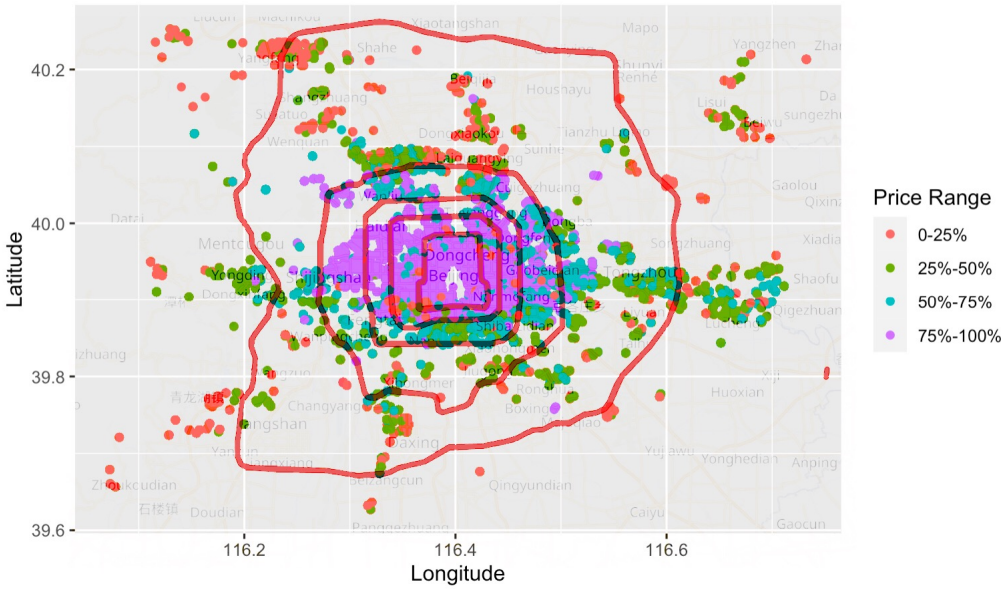

Figure 2: Housing Price Map of Beijing with the Main Rings of Roads (Red Lines)

Comparing the Figure 2, Figure 3Santarelli (2021) shows the housing price map of California and the housing price near to the sea has higher housing price than the inland area do. Similarly, counties like Los Angeles, Santa Barbara, Santa Clara

Table 1: Price Distribution within the No. Ring Roads in Beijing

No. Ring Roads	**Median Price Per Square**	**Increasing Rate**
Beyond 5th Ring Road	28048.75	-
Between 4th and 5th Ring Roads	38726.00	38.06%
Between 3rd and 4th Ring Roads	53788.00	38.89%
Within 3rd Ring Road	156250.00	190.49%

also has more entertainment and business centers like the area within the 3rd ring road. However, Beijing's housing price distribution is mainly determined by the rings of roads, and the districts do not play a significant roles but California's price distribution is mainly determined by counties.

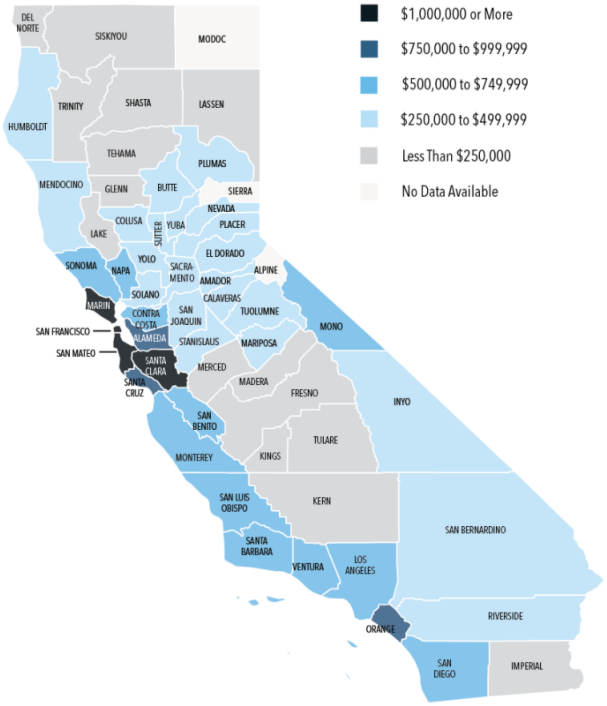

Figure 3: Housing Price are Highest in California's Coastal Urban Area

In the Figure 4, we calculate the Pearson correlation, the statistical relationship or association between two continuous variables. We can see the longitude, latitude, year, renovation conditions, and number of kitchens, living rooms have higher correlation with the housing price per square. We will include these variables in our future models. Additionally, there are some other interesting relations. For example, the building year and the renovation, conditions, the number of elevators and the building structure, the number of bathrooms and the square area of the

house have higher correlation. We will try to construct co-predictor in multi-linear model.

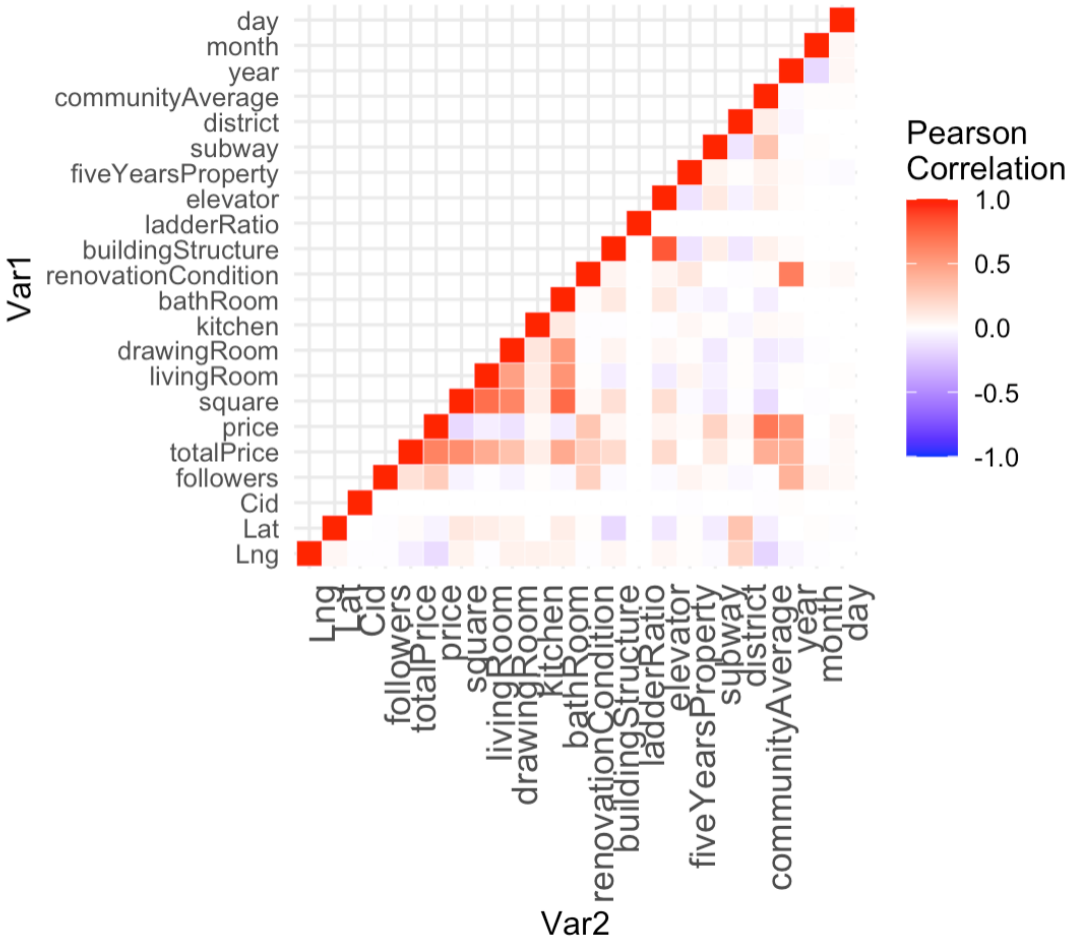

Figure 4: Local Attributes Show Higher Correlation with the Housing Price

Besides the Pearson Correlation on continuous variables, we also plot the feature importance in Figure 5. As we can see, the area where the property is located (districts) has the greatest impact on the real estate market, and its importance is close to 0.6. The second important feature is the housing area. By our exploratory analysis in Figure 6, we can see Xicheng, DongCheng, HaiDian have the highest housing price and these three districts near to the center and have many popular schools, IT companies, and entertainment centers.

After looking at the feature importance, we take a brief look at some important continuous variables and categorical variable, district. We can see the range of price is relatively large than community average. This indicates the distribution of the number of communities in Beijing is pretty average, but the price difference is very high.

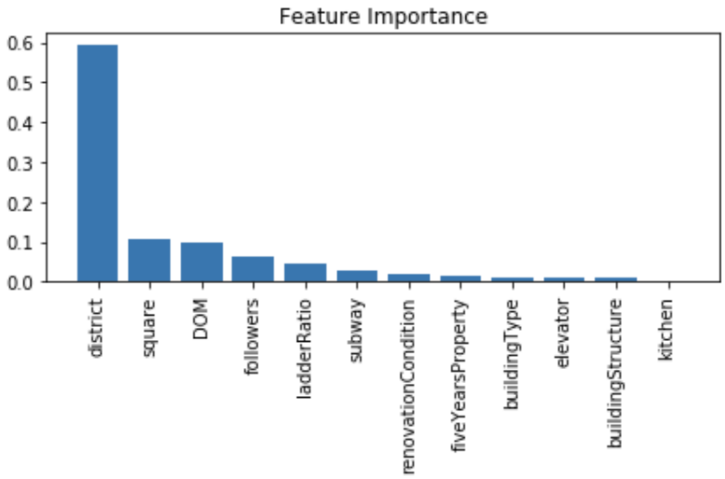

Figure 5: Feature District Shows Highest Importance to the Housing Price

Table 2: Summary Statistics for Important Continuous Variables in the Dataset

Variable	**Minimum**	**Median**	**Mean**	**Maximum**
Square (Area)	7.37	74.29	83.28	1745.50
Renovation Condition	1.00	3.00	2.607	4.00
Community Average	10847.00	59015.00	63684	183109
Year	2002	2015	2015	2018
Total Price	0.1	294.0	349.2	18130.0
Price Per Square	1	38726	43495	156250

From Figure 6, we can see the XiCheng (West City), DongCheng (East City) have higher price than others, since these cities have more political functions. Also, HaiDian have the third highest housing price, because HaiDian own the highest number of schools and technological companies. Many parents desire to own a house in HaiDian so that their children can have a better environment to study and work. Thus, the higher demand of buying a house and the lower supply restricted to the area of HaiDian make the price much higher than other districts.

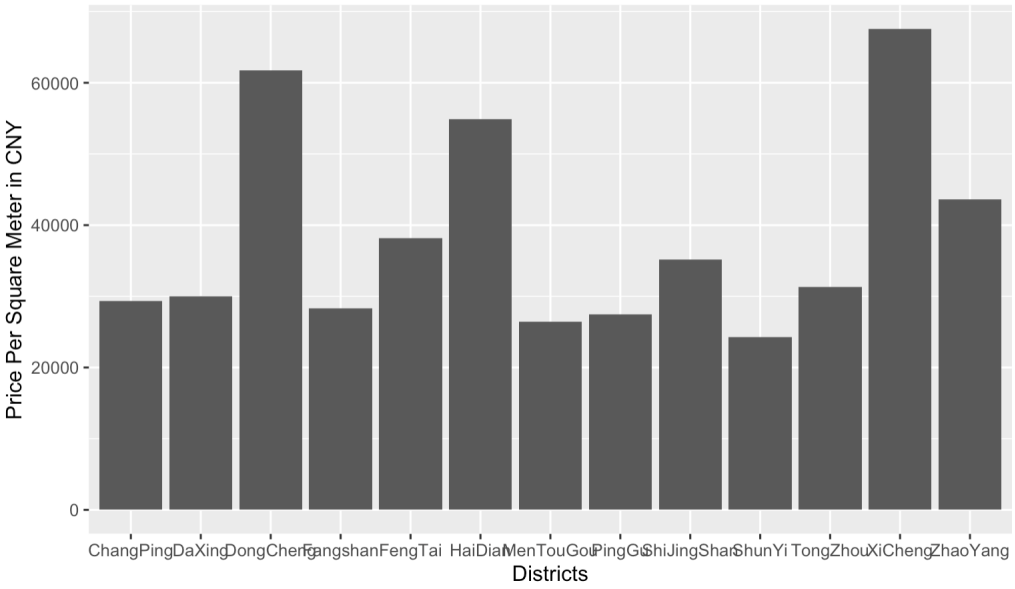

Figure 6: XiCheng, DongCheng, HaiDian Have Relatively Higher Mean Housing Price in Beijing

4 Method

For the method, we use prediction and classification for our variable "price", which is the price in CNY per square meters for each house transaction. Before we start modeling our dataset, we first focus on the specific variable "district", since "district" has the largest feature importance in our exploratory data analysis and it is an important local attributes. To explore how different districts will impact Beijing's real estate market, we use box-plot and price distribution map.

For the model construction part, we transform our longitude and latitude into the distance from the center by using Haversine formula:

$$d = 2r \arcsin \sqrt{hav(la_2 - la_1) + \cos(la_1)\cos(la_2)hav(ln_2 - ln_1)}$$

where la_2, ln_2 denotes the latitude and longitude for each observation and la_1 and ln_2 are the center latitude(39.9°) and longitude(119.4°) and $hav(x) = \sin^2(x/2)$.

Additionally, we transform the subway (if there is subway nearby), elevator (if the house has a elevator), five-year property, building structure, building type, renovation condition into the class character. We also split the trade time into three columns with year, month, and day. After doing these, we split out whole dataset into training set (80%) and test set (20%). Then we do the multi-linear regression, ridge regression, and lass regression for the prediction of price. We

output the coefficients of these three models and compare the R-square and RMSE for both training set and test set.

- Subway: the house has subway nearby (1) or does not have subway nearby(0).
- Elevator: the house own elevator(1) or does not own elevator(0).
- Building Type: including tower(1), bungalow(2), combination of plate and tower(3), plate(4).
- Building Structure: unknown(1), mixed(2), brick and wood(3), brick and concrete(4),steel(5) and steel-concrete composite (6).
- Renovation Condition: including other(1), rough(2), Simplicity(3), hard-cover(4).
- five-year property: the owner has the property for less than 5 year (1) or does not have the property for less than 5 year (0).

Furthermore, after applying the same transformation for the variables listed above, we do the classification for the price, and label "high" for the price greater than the median and label "low" for the price smaller than the median. Then we do the logistic regression for price to see whether logistic regression has a rosy model performance.

5 Results

5.1 Price Distribution by Districts

The 13 districts including Xicheng, Dongcheng, Haidian, etc. have nearly 0.6 feature importance on our response variable price (CNY) per square meters. We output the box-plot and map the distribution for each district. 1-13 corresponds to "DongCheng", "FengTai", "TongZhou", "DaXing", "Fangshan", "ChangPing", "ZhaoYang", "HaiDian", "ShiJingShan", "XiCheng", "PingGu", "MenTouGou", and "ShunYi".

From the Figure 7, we get the similar results as the exploratory data analysis part got that the first three highest housing price are in the districts of Xicheng(10), Dongcheng(1), and Haidian(8). Also, this makes sense in the distribution map, Xicheng(10) and Dongcheng(1) are in the center of Beijing and Haidian(8) is near

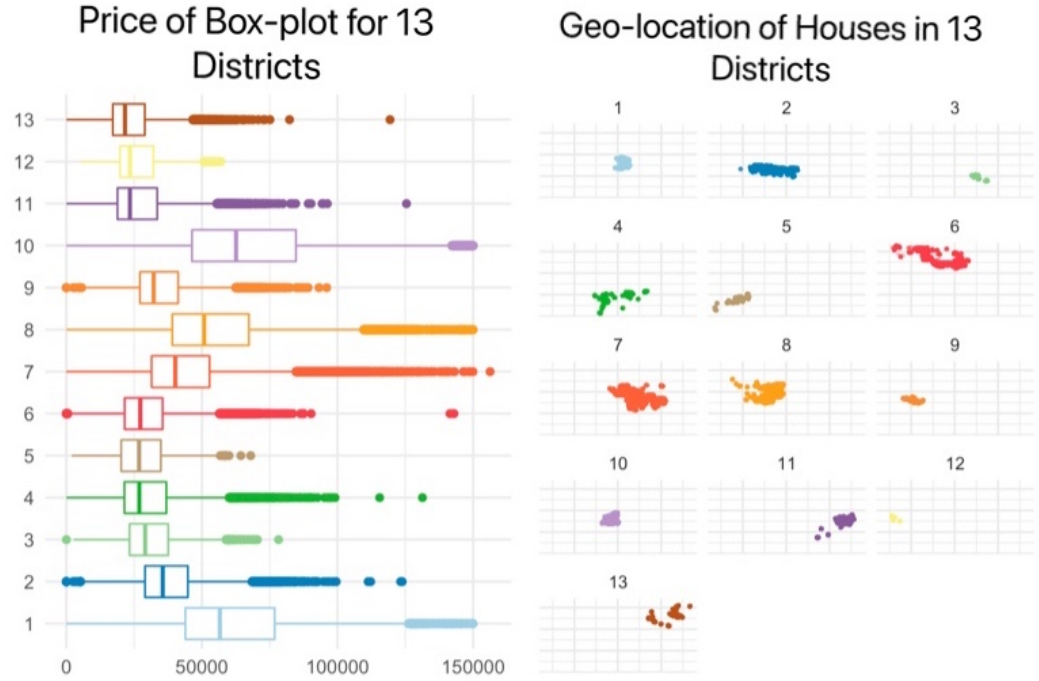

Figure 7: Summary Statistics of the Price in 13 Districts and the Distribution of the Houses depending on the 13 Districts

to these two cities. In Beijing, the closer the houses to the center is, the higher price that the house will be. Another thing worth noting is that Zhaoyang(7) has relatively low housing price comparing to other districts, but its location is near to the center. If we examine Zhaoyang(7)'s box-plot, we can see that Zhaoyang(7) has many outliers that raise the price very higher. This is because Zhaoyang(7) had not been explored in 10 years ago, so the price in early era is not very high. The government recently focuses on the construction of Zhaoyang(7) and this raises the popularity of this district. The demand of living in Zhaoyang(7) district is increasing this year and the housing price is becoming higher. However, technically, this graph is not objective enough, since the number of observations for each district is different from each other. For example, Tongzhou(3) and Shijingshan(9) have relatively smaller number of observations than other districts do.

5.2 Prediction - Linear Regression

Now, after we examine the specifically important attribute district in last section, we focus on the model building and model selection. The first model we apply is

linear regression, or precisely multi-linear regression. After the data transformation and the the random division of training and test set, we attempt to find the linear relationship between price (CNY) per square meters.

From Figure 8, we can see each coefficients are not very large. We also examine the 95% confidence interval for each coefficient and find the interval is still remaining small. Thus, the multi-linear regression model is not over-fitting and we may not consider regularization in ridge and lasso regression. Additionally, the p-value for each variable is smaller than 5% significance level except the variable ladder ratio, thus we will include the significant variables in our final model. As we know, we use the family Gaussian in linear regression.

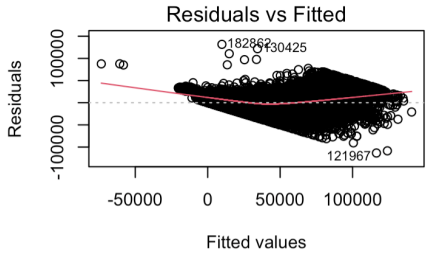

Figure 8: Observations Would Fall Close on the Fitted Regression Line

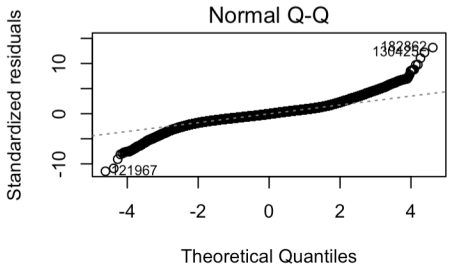

Figure 9: Heavy-Tailed Normal Q-Q plot

From Figure 9 and Figure 10, we observe that the distribution is heavy-tailed, we might agree that the distribution is normal. However, one objection might be that we notice that each end of the distribution seems to extend much further than what we usually see in a normal distribution. Also, the residual line is approximating to zero, this may indicate our multi-linear model works strong in this case. Besides plotting the residual VS fitted values and the normal Q-Q plot, we also calculate the $R^2 = 0.79$ and $RMSE = 9961.55$ for training set and $R^2 = 0.79$ and $RMSE = 10043.79$ for test set.

5.3 Prediction - Ridge Regression

By using the same data processing and same random split of training set and test set, we apply the ridge regression model for our dataset. The ridge regression is an extension of linear regression we just did in last section. The loss function is modified to minimize the complexity of the model by adding a penalty parameter that is equivalent to the square of the magnitude of the coefficients. The summary of the ridge regression is:

```
           Length Class  Mode
lambda     51     -none- numeric
cvm        51     -none- numeric
cvsd       51     -none- numeric
cvup       51     -none- numeric
cvlo       51     -none- numeric
nzero      51     -none- numeric
call        7     -none- call
name        1     -none- character
glmnet.fit 12     elnet  list
lambda.min  1     -none- numeric
lambda.1se  1     -none- numeric
```

Figure 10: The Summary Output of Ridge Regression

To predict our test set and calculate the R^2 and $RMSE$, for ridge regression, we set $\alpha = 0$ and calculate the optimal $\lambda = 100$ and note that the objective function for the Gaussian family in linear is

$$min(\frac{1}{2N}\sum N_{i=1}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda[(1-\alpha)||\beta||_2/2 + \alpha||\beta||_1])$$

Junyang Qian (2021)

We calculate $R^2 = 0.7875102$ and $RMSE = 9962.713$ for training set and $R^2 = 0.7870468$ and $RMSE = 10045.5$ for test set.

5.4 Prediction - Lasso Regression

By using the same data processing and same random split of training set and test set, we apply the lasso regression model for our dataset. Lasso regression is also a modification of linear regression in our 5.2 section. In lasso, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients. In this section, we set $\alpha = 1$ and calculate the optimal $\lambda = 0.001$. The summary of the ridge regression is:

Additionally, We calculate $R^2 = 0.7875598$ and $RMSE = 9961.551$ for training set and $R^2 = 0.7871211$ and $RMSE = 10043.74$ for test set. The Table 3 shows the statistics for all there regression models:

We can conclude that by comparing R^2, the ridge regression model works best to predict the price of house per square meters with $\lambda = 100$.

Additionally, Table 5 is the table of coefficients for all the three regression

```
          Length Class     Mode
a0         1     -none-    numeric
beta      27     dgCMatrix S4
df         1     -none-    numeric
dim        2     -none-    numeric
lambda     1     -none-    numeric
dev.ratio  1     -none-    numeric
nulldev    1     -none-    numeric
npasses    1     -none-    numeric
jerr       1     -none-    numeric
offset     1     -none-    logical
call       6     -none-    call
nobs       1     -none-    numeric
```

Figure 11: The Summary Output of Lasso Regression

Table 3: R^2 and $RMSE$ for Linear, Ridge, Lasso Regression

Models	R^2**(Training)**	$RMSE$**(Training)**	R^2**(Test)**	$RMSE$**(Test)**
Linear Regression	0.79	9997	0.79	10074.32
Ridge Regression	0.7859978	9998.105	0.7857655	10075.67
Lasso Regression	0.786045	9997.002	0.7858254	10074.26

models:

For these variables, only ladder ration has the p-value greater than 0.05. Thus variables except ladder ratio are important.

5.5 Classification - Logistic Regression

As what we state in our method, we also do classification to the housing price. We find the median is 38726 CNY per square meters. We simply classify our price into 1 if it is greater than the median or 0 if it is smaller than the median. Then we do the logistic regression by using the same variables in the prediction model. To add the penalized parameter and make our model more accurate, we also did lasso logistic regression by changing $\alpha = 1$ and calculating the optimal $\lambda = 0.0001$. The training error and test errors are listed in Table 4:

As we can see the lasso logistic regression model has less error than the logistic regression model, though the difference of errors are not obvious. If we do classification for our price, both the lasso and the logistic regression models work well.

Table 4: Coefficients for Linear, Ridge, Lasso Regressions

Coefficients	**Linear**	**Ridge**	**Lasso**
(Intercept)	-1.681605e+07	-1.658331e+07	-1.676089e+07
followers	2.692086e+01	2.783541e+01	2.796686e+01
square	-5.929676e+01	-5.825227e+01	-6.117595e+01
livingRoom	3.409404e+02	3.269525e+02	3.419334e+02
drawingRoom	1.275849e+03	1.231140e+03	1.217064e+03
kitchen	-1.181711e+03	-1.071610e+03	-9.847855e+02
bathRoom	1.181631e+03	1.152207e+03	1.196288e+03
buildingStructure2	1.235719e+03	4.713545e+02	2.543626e+02
buildingStructure3	1.028892e+04	1.126276e+04	1.064662e+04
buildingStructure4	8.430463e+02	8.611135e+01	4.864424e+01
buildingStructure5	1.682026e+03	1.504994e+03	1.282391e+03
buildingStructure6	7.181526e+02	-4.298444e+01	-9.333178e+00
buildingType2	1.290956e+04	1.671247e+04	1.650171e+04
buildingType3	-2.077803e+02	-7.237249e+01	-3.679441e+02
buildingType4	6.493520e+02	7.949239e+02	2.655995e+02
buildingTypeunknown	2.888240e+02	8.529756e+02	-4.099595e+02
renovationCondition2	-8.728668e+03	-8.090981e+03	-8.466684e+03
renovationCondition3	-6.921065e+03	-6.466957e+03	-6.652301e+03
renovationCondition4	-6.061969e+03	-5.628512e+03	-5.973666e+03
ladderRatio	-4.487337e-04	9.295185e-04	1.008777e-03
elevator1	3.375534e+02	4.994532e+02	2.319134e+02
fiveYearsProperty1	-1.521922e+03	-1.470359e+03	-1.293939e+03
subway1	5.837647e+02	6.912291e+02	9.741516e+02
communityAverage	6.275501e-01	6.252914e-01	6.582793e-01
year	8.313117e+03	8.234468e+03	8.320758e+03
month	5.257648e+02	5.256814e+02	5.325678e+02
day	1.832618e+01	2.302405e+01	2.342390e+01
distance	-1.617088e+02	-1.615161e+02	-1.634161e+02

Table 5: Training Error and Test Error for Logistic and Lasso Regression

Models	**Training Error**	**Test Error**
Logistic Regression	0.1145734	0.1150741
Lasso Logistic Regression	0.1145655	0.1150584

6 Conclusion

Along with the local attributes such as longitude, latitude, districts, the housing price in Beijing is distributed by the rings of road. Usually 1st to 3rd rings of road remain in the similar price range. Besides the rings of road determined by longitude

and latitude, the housing price in Beijing have a large standard deviation between districts and districts with long history and near to the center usually have higher housing price. In addition, longitude, latitude, year, renovation conditions, and number of kitchens, living rooms have higher correlation with the housing price per square. After exploring these, we also construct the models for prediction and classification. The best model to predict the price is the ridge regression with $R^2 = 0.786$ and the best model to classify the price is the lasso logistic regression with the test error 0.1150584.

For future work, we may look at the prediction in a different way. For example, we will try to construct prediction model for particular house or for particular year. Besides these attributes, we may add some other variables like the number of schools nearby, GDP, average price nearby, bus transportation center.

References

Junyang Qian, K. T. (2021). An Introduction to glmnet. *R Studio*.

Qiu, Q. (2018). Housing price of Beijing from 2011 to 2017, fetching from Lianjia.com. *Kaggle*.

Santarelli, M. (2021). California Real Estate Market: Prices, Trends, Forecast 2021. *Norada*.