****

Final Project Report

**Beijing House Pricing Analysis**

ALY 6110 Big Data and Data Management

Yi Yang

Instructor: Daya Rudhramoorthi

02/26/2021

**Introduction**

One of the necessities is the house which will be bought by people one or more in a lifetime. The nontransparent house price in the current Beijing real estate market has raised our attention. By exploring the current dataset from local real estate companies, we help to address the important factors forming the current price. In this paper, we will build up the model to predict the price of houses located in Beijing to solve both customer and real estate companies achieve the equilibrium price.

3 Business questions are:

Tools used: RStudio and Tableau

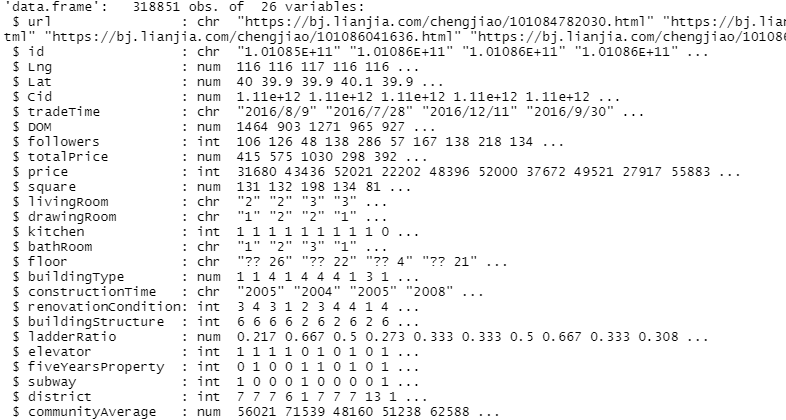
Software language used: R

**Methodology**

***Data Collection***

First of all, we obtained a dataset in the form of csv containing Beijing’s housing price information from a Chinese housing website <https://bj.lianjia.com/ershoufang/>.

After importing the csv file into the RStudio IDE, we now use the str() function to take a peek at the datatypes of all attributes and get to know each of them.



*Figure 1.* Structure of Original Dataset

We can tell that this dataset contains a total of 318,851 observations and 26 attributes. Among it, 8 of the attributes are characters, other 18 attributes are either numeric or integers.

After that, we specifically get to know about each attribute through the following attribute dictionary. The key factor that we are focusing on is the variable: totalPrice, which represents the total price of houses and we will explore how each attribute relates to it.

|  |  |  |
| --- | --- | --- |
| Attribute | Attributes Description | Attribute Value Level |
| url |  |  |
| id |  |  |
| Lng |  |  |
| tradeTime | The time of transaction | Range between 2017-01-01 and 2017-12-31 |
| followers | The number of people follow the transaction | Range between 0 and 1143 |
| totalPrice(Dependent variable) | The total price of houses (Unit 10k dollars, changed from CNY to USD) | Range between 0.15 and 753.68 |
| price | The average price by square (Unit dollars, changed from CNY to USD) | Range between 20.92 and 23071.75 |
| square | the square of house (China official general unit) | Range between 15 and 532.25 |
| bedRoom | the number of bedroom | Range between 0 and 7 |
| drawingRoom | the number of drawing room | Range between 0 and 4 |
| kitchen | the number of kitchens | Range between 0 and 3 |
| bathRoom | the number of bathrooms | Range between 0 and 6 |
| floor | the height of the house | Range between 1 and 63 |
| buildingType | including tower(1) , bungalow(2)，combination of plate and tower(3), plate(4) | Range between 1 and 4 |
| constructionTime | the time of construction | Range between 1950 and 2016 |
| renovationCondition | including other (1), rough(2),Simplicity(3), hardcover(4) | Range between 1 and 4 |
| buildingStructure | including unknow (1), mixed(2), brick and wood(3), brick and concrete(4),steel(5) and steel-concrete composite (6) | Range between 1 and 6 |
| ladderRatio | The proportion between number of residents on the same floor and number of elevators of ladder. It describes how many ladders a resident has on average | Range between 0 and 10009400 |
| elevator | have (1) or not have elevator (0) | 0 or 1 |
| fiveYearsProperty | if the owner has the property for less than 5 years | 0 or 1 |
| subway | have (1) or not have subway (0) | 0 or 1 |
| communityAverage | The average price of the community (Unit dollars, changed from CNY to USD) | Range between 2272 and 24449 |

*Table 1.* Data Dictionary

***Data Cleansing***

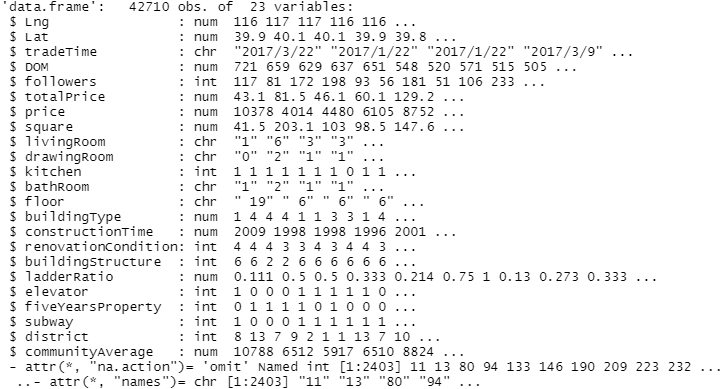
Then, we started preprocessing our dataset in the following steps:

**1. Drop irrelevant columns**: we removed columns of url, id and cid since these columns are not directly related to the totalPrice attribute.

**2.** **Check for missing data**: we used summary() function to check for null values(NA’s), then we noticed there are 6 columns including DOM, buildingType, elevator, fiveYearsProperty, subway, communityAverage that all have NA’s. Among them, 49.54% of the DOM data is missing which could largely impact the final results, so we used the describe() function to take a look at the feature analysis and found out that it is positively skewed. Then, we decided to use the median to impute the DOM data. The percentage of missing values in the other 5 columns are all below 0.63% and not that significant. Therefore, we can exclude these, and we are left with 316448 records and 23 variables.

**3. Data subsetting**: we extracted only the data from the Year 2017, now we are left with a total of 42710 cases.

**4. Data reformatting**: we removed error words in floor. Revalued the totalPrice and price according to the inflation rate and changed the unit of these two variables from CNY to USD based on the google currency exchange rate. Organized the dataset according to transaction time. Our final dataset has 42710 records and 23 attributes.



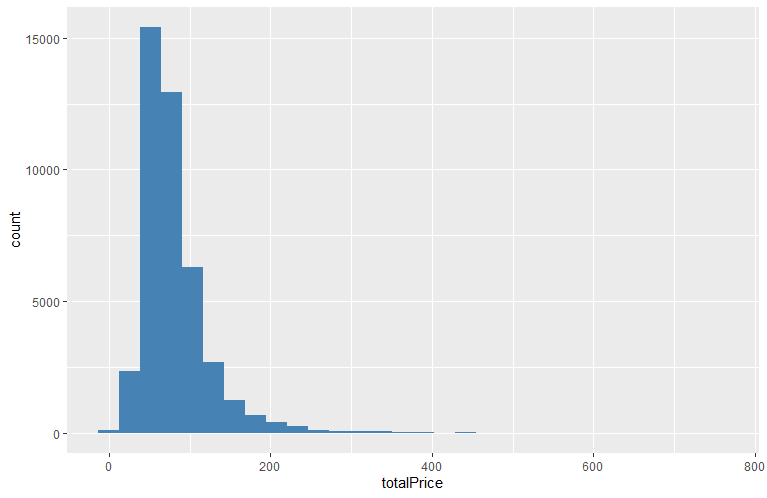
*Figure 2.* Structure of Final Dataset

**5. Outlier detection**: before removing it, we found that there are 1189 outliers compared to the full data with 42710 records. Since our project is mainly focusing on the housing factor, having this number of outliers could make our future modeling go wrong. Therefore, by not deleting any total price, we choose to change all outliers to mean plus two standard deviation. This represents that all outliers are in the highest price, but not getting any outlier penalty.

6. **Dummy variables creation**: to prepare for later regression analysis, we converted all the categorical variables into dummy variables that are binary variables (0 or 1) to represent the subgroup under each attribute.

***Exploratory Data Analysis (EDA)***

First of all, a histogram is plotted to take a look at the distribution of our target variable totalPrice.

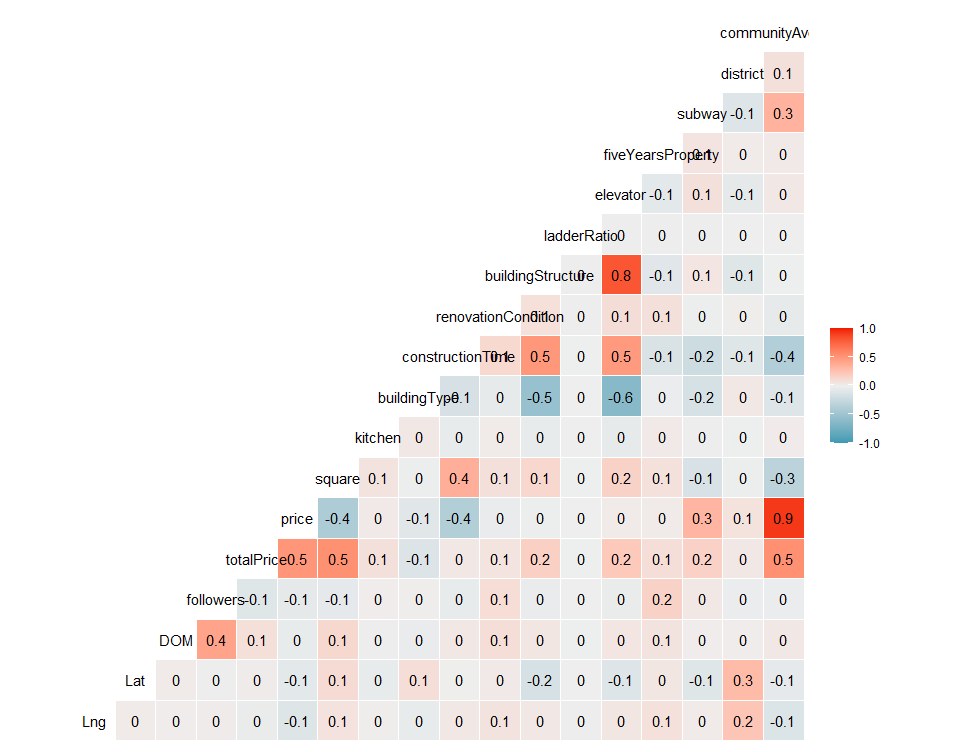


*Figure 1.* Histogram of Total Price

We also made two plots about whether having an elevator or subway or not will affect the total price of houses.

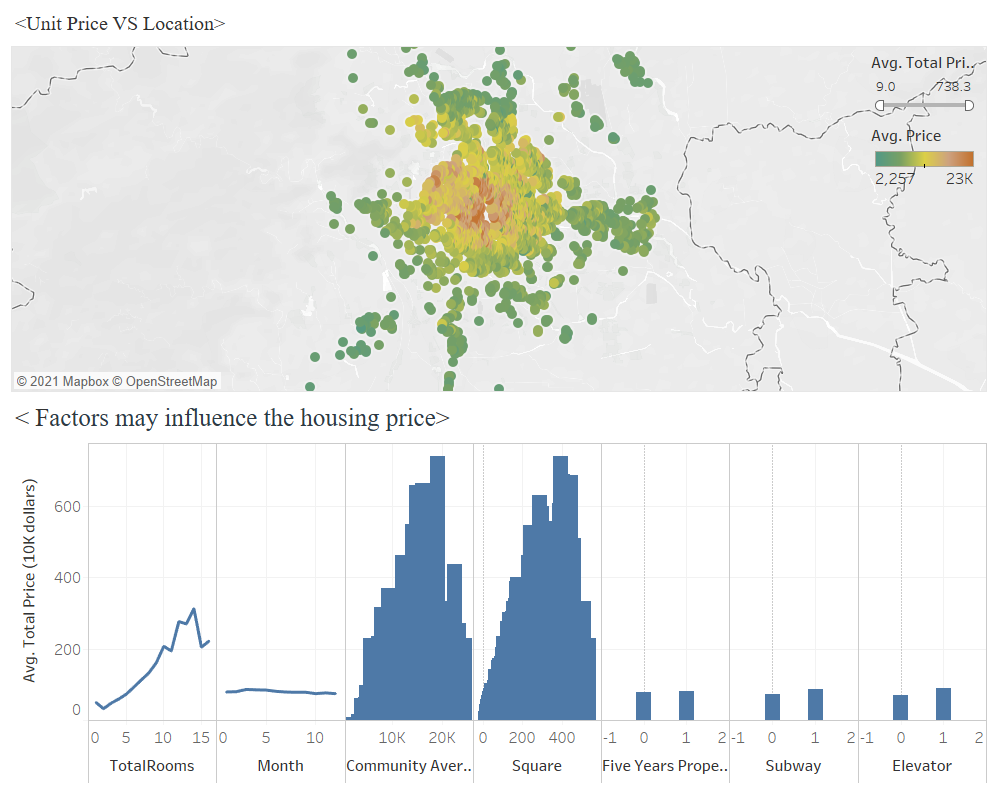
We can observe from the above two plots that there are some differences in total prices on whether an elevator or subway is available. Houses with elevators or close to the subway are a bit higher than the ones without.

From this plot, we can tell that most of the observations fall into the range of around 200k to 900k dollars. When the total price gets higher close to 3000k dollars, the distribution gets almost flat. Then, we used the ggcorr() function to plot the correlation matrix to explore what key factors that affect the total price.



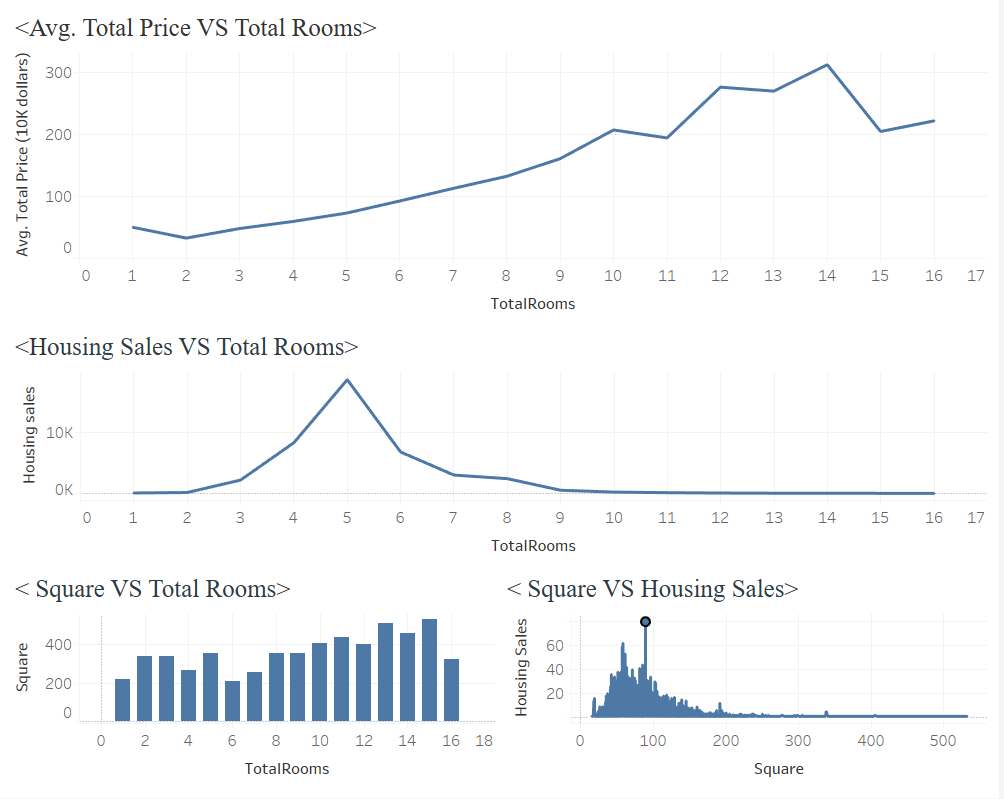
*Figure 4.* Correlation Plot of our dataset

As Figure 4 shows, square and communityAverage are most positively correlated with totalPrice.



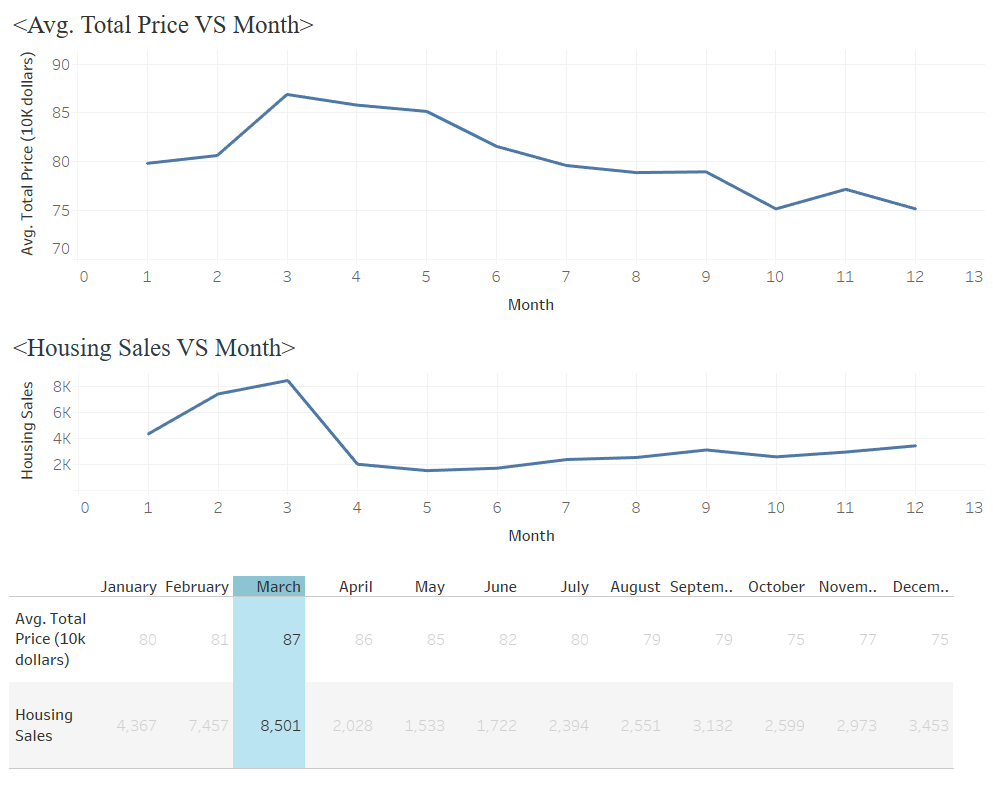
*Figure 17.* Feature & House price

From the Figure 17 dashboard, we can see that the unit price of the downtown area is the highest, as the location becomes closer to the center part of the city, the unit price climbs. Room Types, Community Average, and Square also will influence the overall house price. However, tradeTime, five-year property, subway, and elevator didn’t affect the total house price a lot which may answer the question which factors may influence the housing price.



*Figure 18.* House Types

Furthermore, in order to answer what the most popular house types are, from the Figure 18 dashboard, it is obvious that average total prices will increase as the total number of rooms increases. However, when the room numbers increased to 15, the prices have dropped based on the phenomenon that there are buyers who gather other buyers to buy a house and cut the living places into small rooms. Moreover, square does have a slight influence on housing sales, but the total number of rooms will not be affected. Compared with other house types, houses with 5 rooms and square equal to 89 m2 have the most transactions sales. Giving us the information that five rooms’ houses can be favored by most house buyers in Beijing.



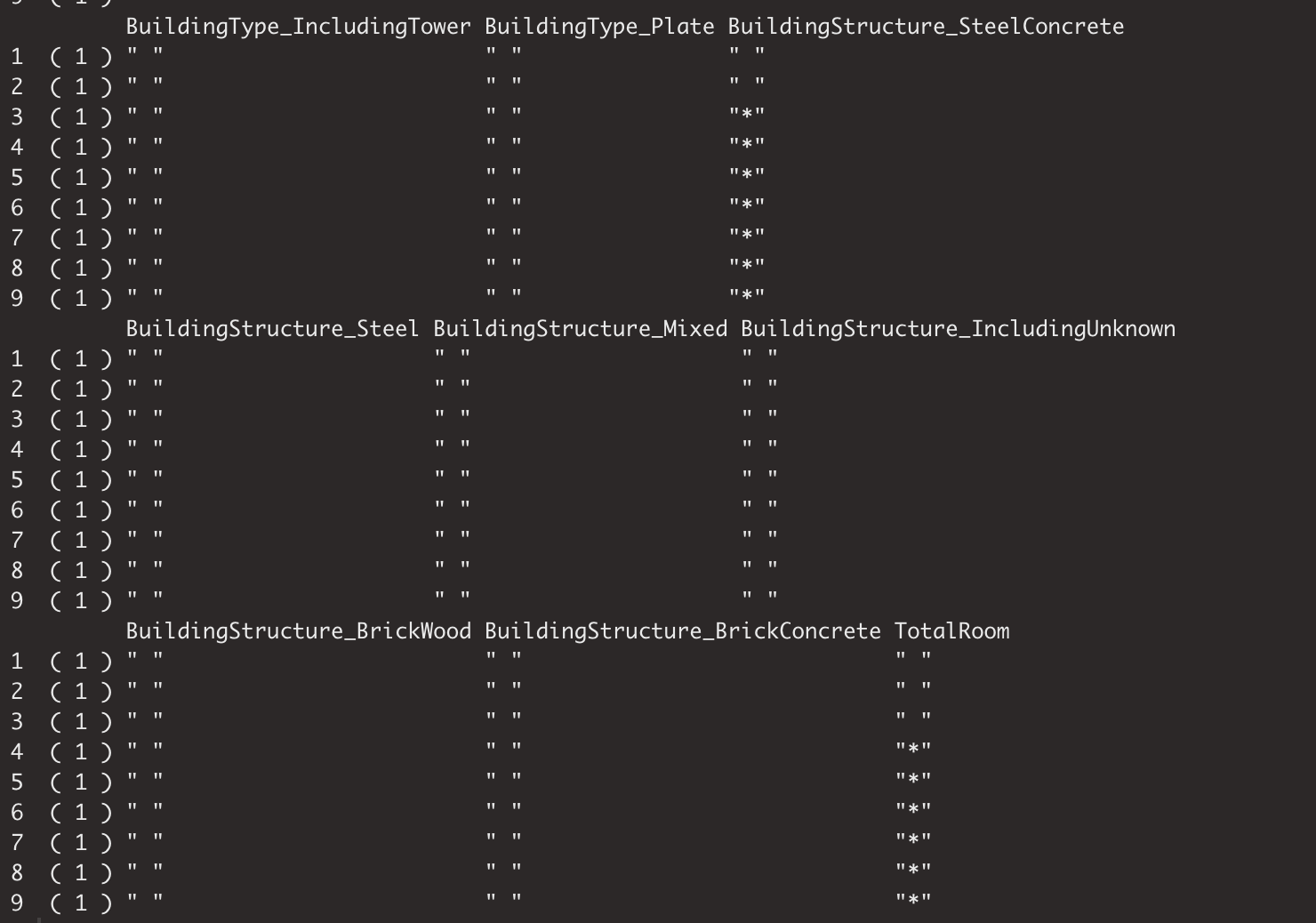
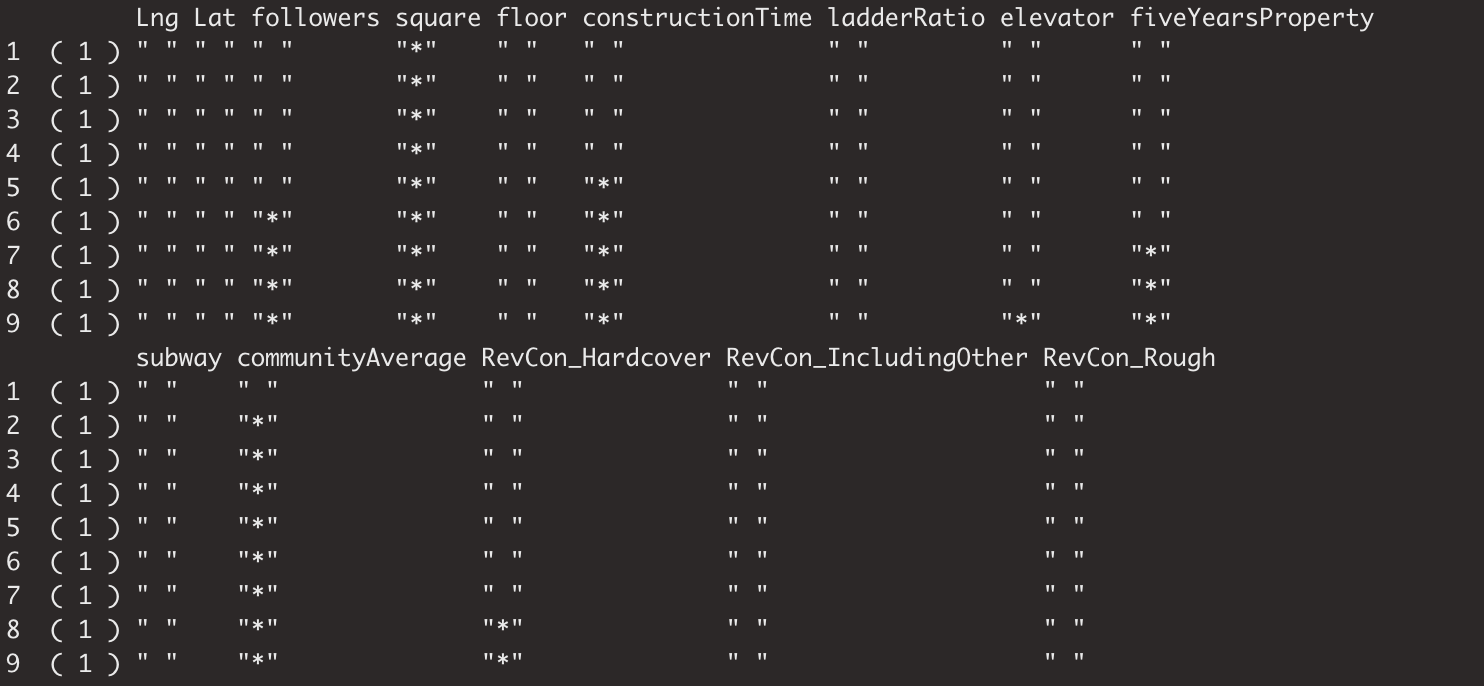
*Figure 19.* House price VS Time

In Figure 19, we can see the highest count of trades happened in March which impact on the average total house price was the highest (around 870k). Hence, people in Beijing may have the highest housing demand in March. It may give Real Estate Companies an insight on March is the best time to sell houses.

***Data Preparation***

To build the prediction model, we first changed variables into dummy variables, including RenovationCondition, BuildingType, and BuildingStructure. Meanwhile, we created a new column, totalRoom. The reason why we created totalRoom is that the variable bathRoom has a p-value of more than 0.05, which indicates that it is not significant; however, we consider bathRoom as a significant variable regarding business insights. More bathrooms mean the house is bigger, which increases the total price for the house. Therefore, we created a totalRoom column, combining Bedroom, drawingRoom, bathRoom, and kitchen. Furthermore, we changed the outlier of totalPrice to the mean plus two standard deviations to prevent the effect from extreme values.

Besides, we used a stepwise method to find the most relevant variables to build the regression model because we found that some variables have higher p-value, which is not significant for the dependent variables. Thus, through using the stepwise method, we can find the variables that are best fit the model. We chose 10 and 15 to check the optimized number of variables, and the result from step.model$bestTune shows that eight variables can be the best fit for the model. In Figure 5, we can see the eight variables, including followers, square, constructionTime, fiveYearsProperty, communityAverage, RevCon\_Hardcover, buildingStructure\_SteelConcrete, and TotalRoom.



*Figure 5.* Stepwise results

***Model Selection***

Now we consider the options of predictive modeling algorithms toward our dataset. Since our outcome variable is numeric and continuous, the first model we consider is linear regression since we want to examine if a set of predictor variables do a good job in predicting totalPrice and what variables are significant predictors of totalPrice. Besides, achieve a linear regression model between totalPrice and other attributes can help us predict the future totalPrice of houses in Beijing given some important factors.

Random Forest (RF) Analysis is another technique we decided to use toward our dataset since it can be used in both classification and regression tasks, more importantly, it offers high accuracy through cross validation and it has the power to handle a large data set with higher dimensionality.

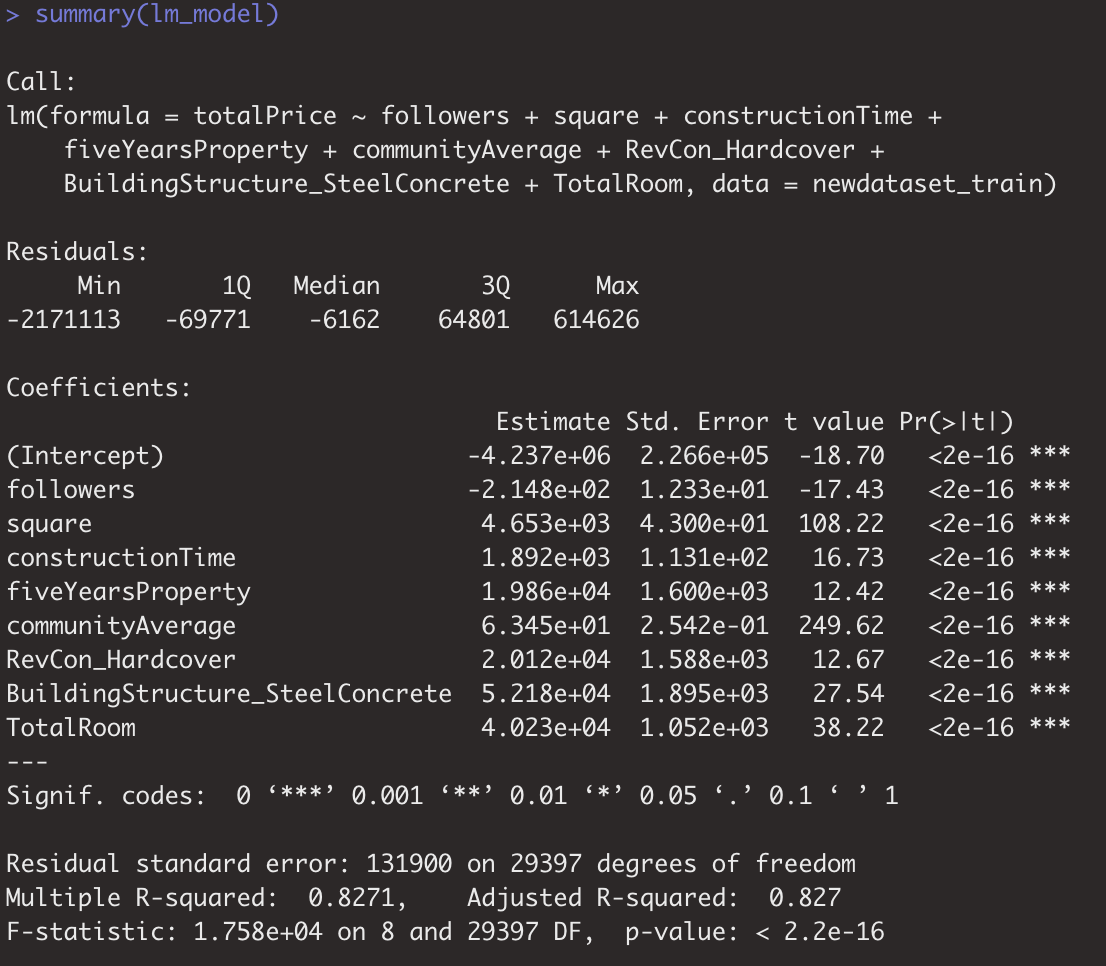
Moreover, we also tried another tree-based algorithm named Gradient Boosting Modeling (GBM) for our dataset since it has been shown that GBM can perform better than RF if parameters are tuned carefully. So, we decided to try it if we can achieve higher accuracy for our model.

***Train/Test Dataset Split***

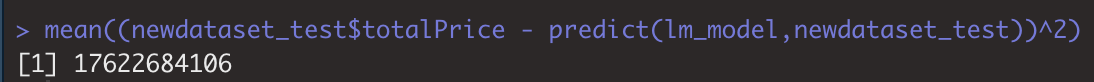
Before building the regression model, we split the dataset into train and test tests because it can minimize the effects of data discrepancies and better understand the model's characteristics. The dataset was separated into the train set and test set, where 70% of the data goes into the train dataset and 30% of the data goes into the test dataset. Therefore, the train dataset has 29406 observations and the test set has 12603 observations.

***Linear Regression Modeling***

Through using the lm() function to create the linear regression model, then we created a prediction model. The summary result for the linear regression model is showing in Figure 6. Then we also calculated the mean square error for the linear model. Figure 6 and Figure 7 indicate that the R-squared is 0.8271 and the MSE is 17622684106.



*Figure 6.* Linear regression model result



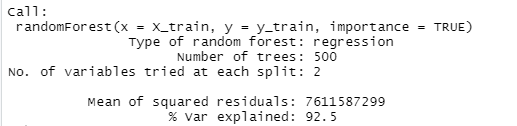
*Figure 7.* MSE for linear regression model

***Random Forest Regression Modeling***

After generating the linear regression model, we further used the random forest algorithm to predict our predictor variable.

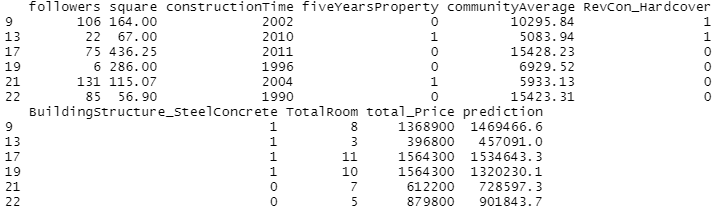
A Random Forest is one type of Decision Tree algorithm and it is capable of performing both regression and classification tasks. It works by creating multiple decision trees and then combining the output generated by each of the decision trees to predict an outcome variable. It helps reduce the correlation between the decision trees and removes the bias that decision trees might introduce in the system.

Below is the Random Forest model that we built based on our dependent variable totalPrice and 8 variables that we selected based on the above Stepwise selection.



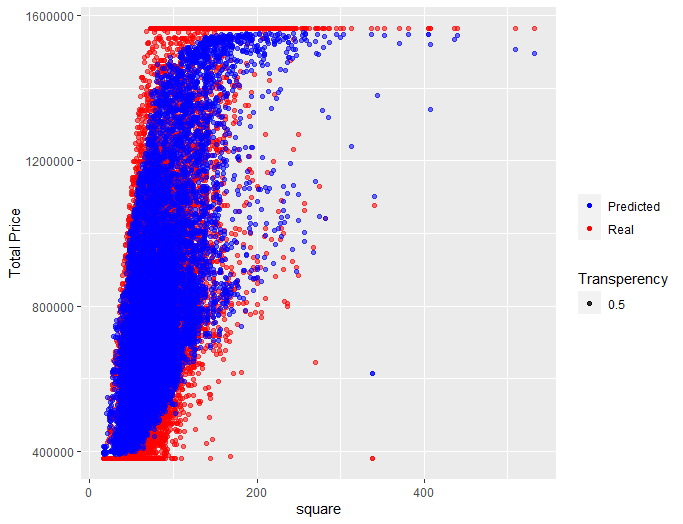
*Figure 8.* Random Forest Model with default parameters

We can observe that the accuracy of our model based on train data is pretty good. Then, we can predict values for the test data. In order to compare the predicted value with the actual values in the test data and analyze the accuracy of the model, we will show it both in the form of the table and plot the total price and square value. And below is the result table.



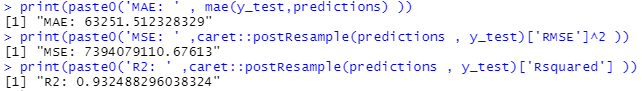
*Figure 9.* Table of Predicted values and Real values

To more visualize the comparison between predicted and real values, we used the ggplot() function to draw the scatterplot below.

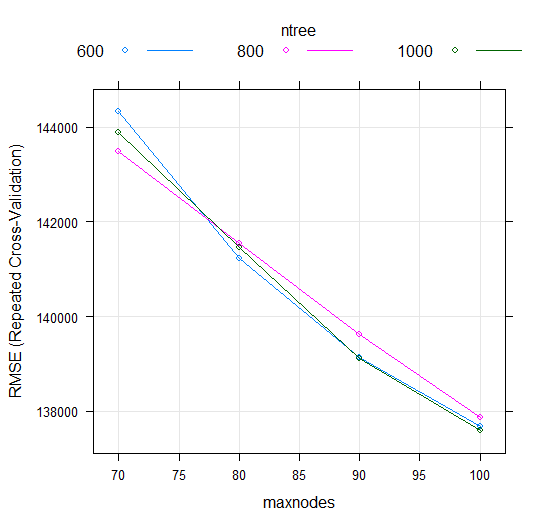


*Figure 10.* Scatter plot between Predicted values and Real Values

We can tell from Figure 10 that the predicted prices (blue scatters) coincide pretty well with the real ones (red scatters). In order to estimate our model more precisely, we will look at Mean absolute error (MAE), Mean squared error (MSE), and R-squared scores.



*Figure 11.* MAE, MSE, and R-squared value of Random Forest Model

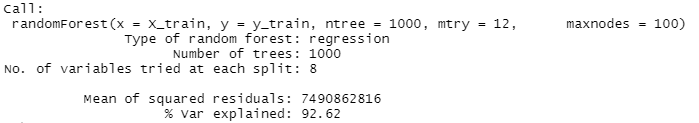
After obtaining high error values (MAE and MSE), we can try to improve the predictive power of the model by tuning the hyperparameters of the algorithm. In order to tune the parameters ntrees (number of trees in the forest), mtry(number of variables split at each node), and maxnodes (maximum number of terminal nodes trees in the forest can have), we will need to build a custom Random Forest model to obtain the best set of parameters for our model and compare the output for various combinations of the parameters. The plot shows how the model’s performance develops with different variations of the parameters. There seems not much difference among options of ntree.

*Figure 12.* Plot of maxnodes vs RMSE



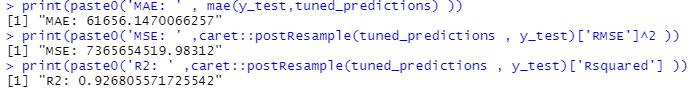
*Figure 13.* Best tune parameters generated

Based on the graphs above, for values maxnodes: 100 and ntree: 1000, the model seems to perform best. We can now use these parameters in the final model.



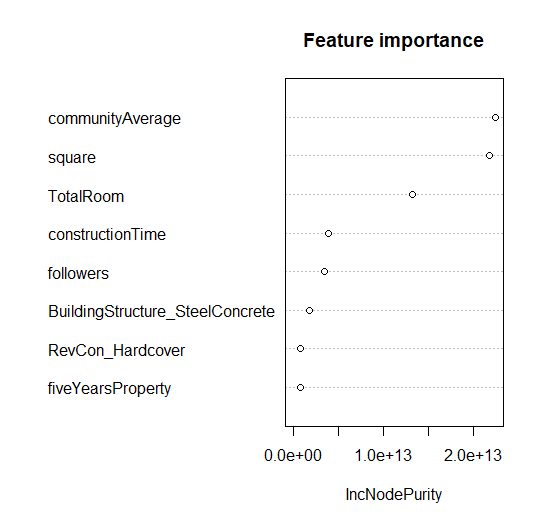
*Figure 14.* Tuned Random Forest Model

Compared to the default model we built, the accuracy improved just a bit after tuning.



*Figure 15.* MAE, MSE, and R-squared value of Random Forest Model

Also, the error rate values got decreased a little. Besides, we made the variable importance plot below.



*Figure 16.* Variable Importance Plot

Based on Figure 16, we can observe that communityAverage and square are the most important variables, which matches the result of our correlation plot.

**Analysis**

Interpretation:

To conclude, we made a comparison table listing three models’ R-squared values and running time.

|  |  |  |  |
| --- | --- | --- | --- |
| **3 Predictive Models** | R2 | MSE | Running Time |
| Linear Regression Modeling | 0.8271 | 1.7e+10 | 1 minute |
| Random Forest (Regression) Modeling | 0.9268 | 0.7e+10 | > 5 minutes |
| Gradient Boosting Modeling(GBM) |  |  |  |

*Table 2.* Comparison between 3 predictive modeling methods

In terms of this dataset, Random forest seems to be a better option for predictive analysis compared to other two models since it has the highest accuracy rate.

**Conclusion**

In this paper, we have discovered the following patterns. First, from the variable selection, we choose the most relevant houses’ attributes which can give companies the view on the factors of houses needed to be aware to get quick sales. Second, from the model selection, we chose the random forest model as the optimal model since it has the advantage on high R squared value and low MSE value. Although it takes time to process, accuracy is more important in the housing market. In the last part, the dashboard helps to locate the housing market pattern giving companies which month of the year and how many rooms are able to achieve the best sales.