BEST OF ZILLOW

NYC Home Sale Price Analysis based on Zillow Data

Team Alpha
Peiying Yu, Jianxing Wan, Ziyuan Wang

April 2019

BEST OF ZILLOW

1. Introduction

Home purchase can be one of the most important decisions in one's life. An estimated of home value would be helpful either before or after owning a home. New York City's real estate prices deter many people, and buyers can only guess at the factors which influence these prices, such as geographical location, crime safety, or transportation convenience. In this project, we researched the potential factors that might affect home prices with information obtained from Zillow and with the tools we learned in the class.

The goals of our project are to identify significant inherent and external features that affect the home price in NYC, and to build and choose the best predictive models for home values based on different estate features.

2. Dataset

Data Acquisition

We use three datasets for our project. First, we built a web scraper with BeautifulSoup to directly scrape features and prices of homes for sale in NYC from Zillow official website [1]. After iterating through 12 neighborhoods and 53 zip codes in Manhattan and Long Island City, we obtained information for 7,665 homes in NYC with the scraper. The key features we scraped are: number of bedrooms and bathrooms, year-built, price, house size, zip code, latitude, longitude, neighborhood, overview description of house, and address. We refer to those features as home's inherent features and this dataset as Major Data Source in analysis later.

In order to analyze the relationship between home price and transportation convenience, we obtain the second dataset "NYC Transit Subway Entrance And Exit Data" from New York state official website [2]. This open dataset includes features of station name, subway route, latitude and longitude of entrance and exit, for 1,868 subway entrance and exits in NYC. We refer to this dataset as Auxiliary Data Source I.

Finally, in order to analyze the relationship between home price and area safety, we obtain the third dataset "Crime Complaint Data" from NYPD [3]. This open dataset include features of Crime Complaint Number, Crime Complaint Date, Complaint Latitude and Longitude, for 114,673 crime complaints in NYC. We refer to this dataset as Auxiliary Data Source II.

Data Processing

a. As discussed in the Data Acquisition section, we divide our data set into 12 neighborhoods based on NYC government neighborhood definition [4]: Central Harlem, Chelsea and Clinton, Gramercy Park and Murray Hill, etc.
b. In order to obtain the factor 'distance to subway', we use the longitude and latitude coordinate of each home from Major Data Source and those of each subway station from Auxiliary Data Source I to calculate the distance from one home to its nearest subway station. In order to convert distance between longitude and latitude coordinate to distance in kilometer, we used Haversine formula [5]:

$$d = 2r \arcsin\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + \cos\phi_1 \cos\phi_2 \sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)},$$

where point 1 (ϕ_1, λ_1) is the latitude-longitude coordinate of home, point 2 (ϕ_2, λ_2) is the latitude-longitude coordinate of each subway station entrance/exit, d is the distance in km between point 1 and point 2. We use the radius of Earth for r ($r_{Earth} = 6371$ km).

With this formula, we don't need to use Google GeoAPI to accomplish the conversion, and could save a lot of time.

c. Utilizing the latitude and longitude data of crime complaints from Auxiliary Data Source II, we calculate the total number of crime complaints in each zip code area, and then merge the crime complaint data into data frame of Major Data Source by zip code as one of our Data Frame feature ('complaint').

d. We added a feature 'price range', which divides home for sale prices into 31 categories. We created each categories with $0.2M increment starting from $200,000, and with $1M increment for house prices greater than $3M.

3. Exploratory Data Analysis (EDA)

After merging different data sources, we get a final Data Frame with diverse characteristics including inherent estate features: 'bed', 'bath', 'year built', 'price', 'address', 'description', 'latitude', 'longitude', 'zip code', 'size', 'Neighborhood' from Zillow, and external estate features: 'distance to subway' and 'complaints'. Then, we analyzed the dataset to find some interesting patterns and provide useful insights.

We divide potential factors into two classes: (1) Inherent factors: features that can directly get from Zillow; (2) External factor: features never appear on Zillow.

I. Price Distribution and Inherent Features Analysis (Based on Major Data Source)

a. We grouped by zip code to calculate the average price in each zip code area and show the price distribution in total Manhattan by heat map (Figure 1). It clearly indicates that Lower Manhattan, Murray Hill and Chelsea and Clinton has higher price by the heat map.

b. We grouped by the 'Neighborhood' feature and found that the location of maximum price interval value was "Greenwich Village and Soho" using BoxPlot (Figure 2).

c. We made a correlation table (Figure 3) of features of this Data Frame and found out the number of bathrooms and beds has strong positive correlation with estate's price.

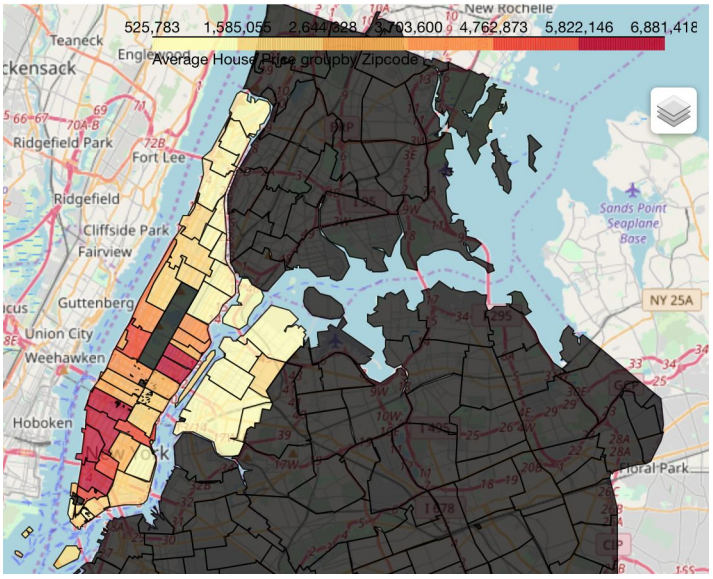


Figure 1: Heatmap of Average price

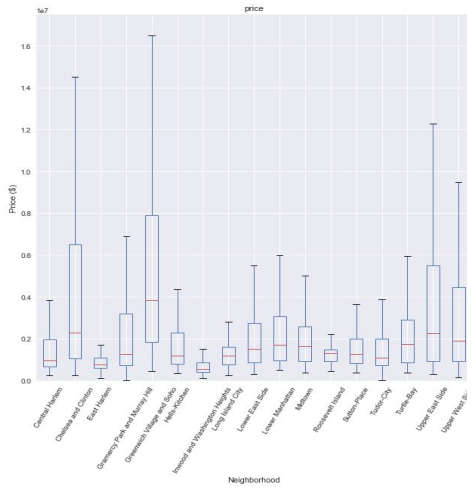


Figure 2: Boxplot of home prices by neighborhood

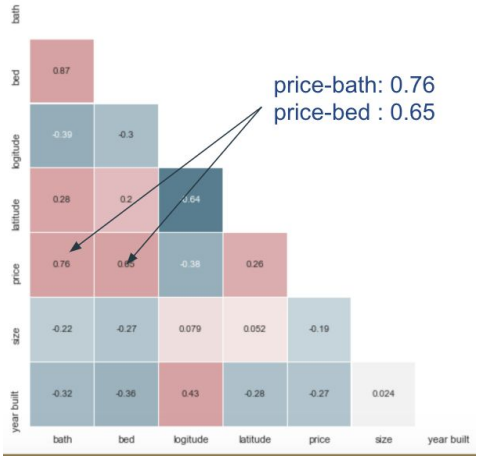


Figure 3: Correlation Plot

d. We also generated Word Clouds from 1,700 house overview descriptions of 4 areas in the "Greenwich Village and Soho" neighborhood (Figure 4). We found that the facilities service, landmark building and scenery may affect house price. However, those are inherent factors of the house to affect price. We hope to get more external factors to affect the price.

Figure 4: Word Clouds

e. Using LDA thematic analysis to the overview descriptions (Figure 5) we listed 30 most salient terms. We found that transportation and safety are two possible key factors. This is consistent with our renting experience in real life.

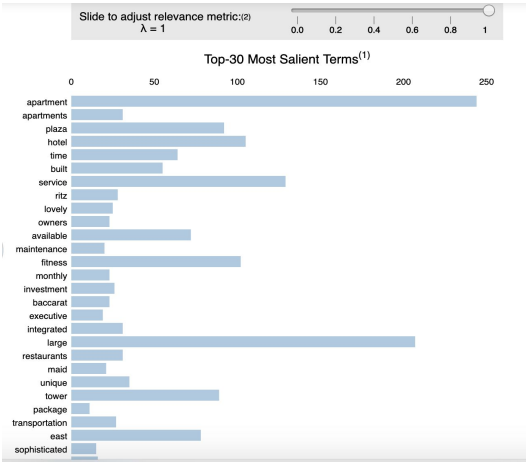


Figure 5: LDA thematic Analysis

II. External Features Analysis (Based on Auxiliary Dataset I &II)

a. After calculating number of crime complaints in each zip code area, we generated a heat map of crime complaints (Figure 6). Comparing to the price distribution in Figure 1, we find

lower Manhattan area with higher house prices has higher crime complaint rates, while the Murray Hill area also has higher house prices but lower crime complaint rate. It seems that there is no correlation between house price and crime complaint rate based on the whole dataset. We also made a correlation table with crime rate feature in data frame, and it also shows there is no obvious correlation (ϱ=-0.041) between price and the crime complaint rate.

Based on the findings above, we grouped the data by zip code area, and find that home price can be either positively or negatively correlated with the crime rate, conditional on zip code or neighborhood (Figure 7). Hence, the external factor crime rate should be considered for regional home pricing.

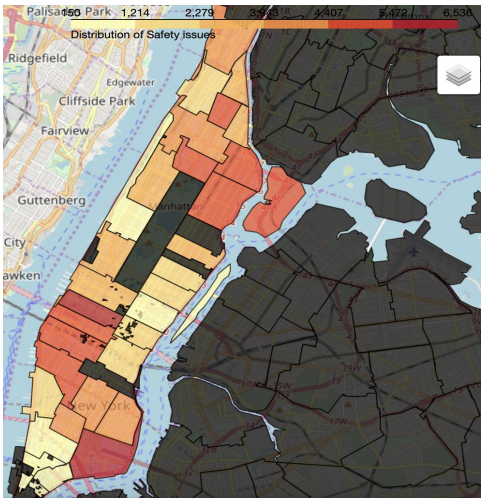


Figure 6: Heatmap of Crime Rate

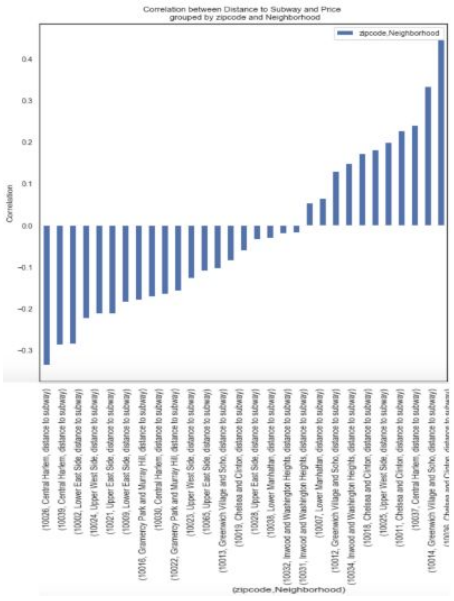


Figure 7: Price correlation with crime rate by zip code

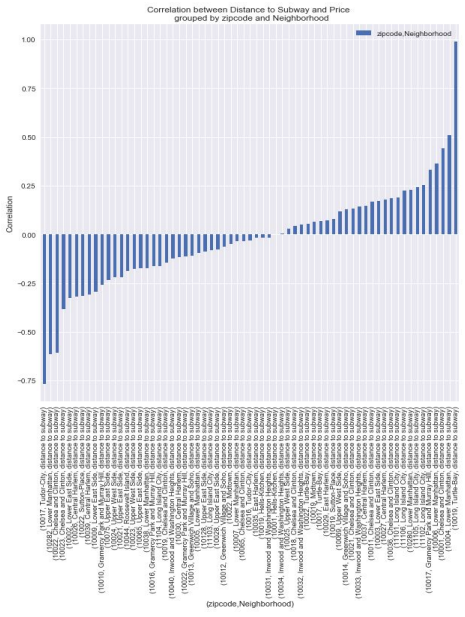


Figure 8: Price correlation with subway distance by zip code

b. Similarly, we initially found no obvious correlation (ϱ=-0.065) between price and the external factor 'distance to closest subway' based on the entire dataset. However, after we grouped the data by zip code area, we found that home price can be either positively or negatively correlated with its distance to subway, conditional on zip code or neighborhood (Figure 8). For some region (Turtle Bay, 10019), the correlation is as high as ϱ= 0.99, while some (Tudor City, 10017) has as low as ϱ= -0.77. Hence, the external factor distance to

subway should be considered for regional home pricing as well.

4. Predictive Analytics

We used three different methods to predict the house price, price range and price for certain region. We ran the following three models first to predict price and price range, and finally on a specific region (zip code 10036 in Chelsea and Clinton) to predict price range. We split the data into train and test in order to conduct model selection for random forest and KNN.

I. Linear Regression

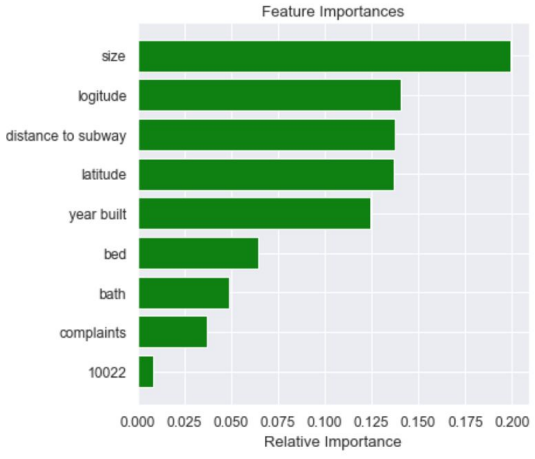
We first run a linear regression model on the initial dataset, but the model didn't fit well. We think the misfitting is caused by outliers. Therefore, instead of predicting the exact price value, we turned to predict the price range to which each house belonged. Based on EDA, some features have strong correlation with zip code. Thus, we tried to fit a linear regression model for each region. MSE and R - squared improved a lot from our original naive model. Significant features are bath, bed, longitude, complaints. We also identified the significant features based on p-values (Figure 9). The significant features are: number of bathrooms and bedrooms, longitude and number of complaints.

	coef	std err	t	P>|t|	[0.025	0.975]
bath	1.5519	0.387	4.012	0.000	0.784	2.320
bed	1.1628	0.468	2.486	0.015	0.234	2.092
latitude	110.8840	271.862	0.408	0.684	-428.980	650.748
logitude	-419.8556	206.948	-2.029	0.045	-830.812	-8.899
size	-3.441e-06	4.19e-06	-0.821	0.414	-1.18e-05	4.88e-06
year built	0.0018	0.015	0.119	0.905	-0.029	0.033
distance to subway	0.2264	3.903	0.058	0.954	-7.525	7.978
complaints	-22.1509	7.487	-2.959	0.004	-37.019	-7.283

Figure 9: Feature Importance from Linear Regression

II. Random Forest

Random forest told us the most important features are size, longitude, distance to subway, latitude, year built, bed and bath (Figure 10). In order to find the optimal random forest model, we trained our model using multiple different depths of trees. According to the MSE on test data, the optimal depth is five (Figure 11). Compared to Linear Regression, Random Forest performed better, as shown by the table above.



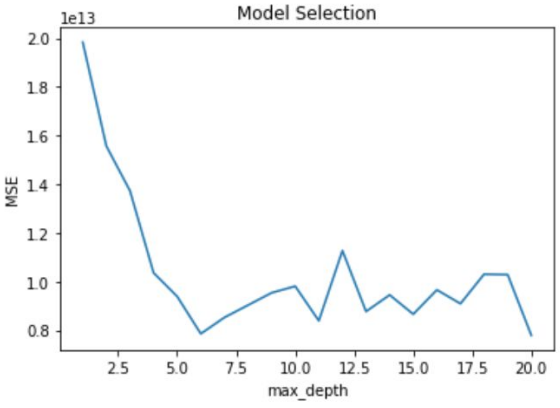


Figure 10: Feature Importance from Random Forest

Figure 11: Optimal Depth = 5 for Random Forest

III. K Nearest Neighbors (KNN)

Since homes with geographical proximity and similar features should have similar prices, we ran KNN models. The features we used in this model are latitude, longitude, bed, bath, distance to subway and complaints. We used these features because they are the significant ones identified in the previous two models. For house price prediction, we use KNN regressor, since price is a continuous variable. For price range prediction we use KNN classifier, since price range is a categorical variable. According to the MSE on test data, the optimal k for KNN regressor is 18 and for KNN classifier is 7. Finally for specific region, we ran KNN classifier for price range prediction with optimal k = 14.

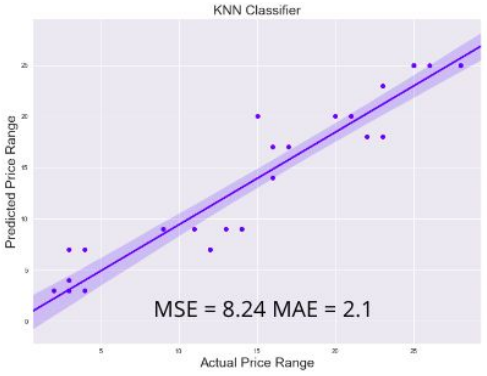


Figure 12: KNN Performance for Regional Price Range

The MSE table (Table 1) shows that prediction based on specific region is better than on the whole NYC dataset. This finding is consistent with the previous correlation analysis in section 3 that factors have more explanatory power on a regional house price .

	MSE on Test Data		
Model	Price	Price Range	Region
Linear Regression	6.7e+13	15.99	12.72
Random Forest	**9e+12**	14.72	13.56
KNN	2.4e+13	**12.06**	**8.24**

Table 1: MSE for 3 Models

5. Conclusion

From exploratory data analysis (EDA), we identified that numbers of bedrooms and bathrooms are inherent factors that have significant impact on home price based on correlation analysis. Crime complaint rate and distance to subway station are external factors that have significant impact on home price in a certain region specified by zip code . In addition, sentiment analysis with word cloud and LDA thematic analysis also shows that transportation and safety are two key factors for home price.

Based on those facts, we ran linear regression, random forest, and KNN models to predict price and price range with key features identified in EDA. We find that prediction based on a specific zip code. The best model based on test data MSE is KNN for regional price range prediction with k = 14.

6. References

1. https://www.zillow.com/homes/for_sale/
2. https://data.ny.gov/Transportation/NYC-Transit-Subway-Entrance-And-Exit-Data/i9wp-a4ja
3. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/
4. https://en.wikipedia.org/wiki/Haversine_formula
5. https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
6. https://github.com/yupeiying1/Zillow-Home-Value-Prediction (GitHub repository for this project)

BEST OF ZILLOW

For any questions regarding this report, please contact:

Peiying Yu at py2244@columbia.edu

Jianxing Wan at jw3693@columbia.edu

Ziyuan Wang at zw2396@columbia.edu