Data Analysis Project – Airbnb Chicago

Introduction

In this project, we explore the Airbnb datasets of Chicago, which contain a list of all Airbnb records in Chicago, Illinois, for each month from April, 2018 to March, 2020 (data of June, 2018 is missing) (*http://insideairbnb.com/get-the-data.html*). In these datasets, there are plenty of variables for every single Airbnb record, including Airbnb names, host names, neighborhood, longitude, latitude, room type, price, total numbers of reviews, review per month, availability, etc.

Looking into the dataset, we have raised several interesting questions, and presented the corresponding visualizations to better understand and answer them. From the visualizations, we also get some fresh insights and will dig into the problems in detail by using some popular data analysis tools, like regression, cross validation and forecasting. For unchronological problems, we only focus on the data of March, 2020 (*http://data.insideairbnb.com/united-states/il/chicago/2020-03-18/data/listings.csv.gz*) since it is the most recent and representative dataset.

For tourists, when they plan to visit somewhere, one thing that they most concern is where to live. To better serve the customers, we want to show clearly where most people choose to live and how expensive it is in a certain area. Furthermore, we will try to analyze the fundamental drivers, for example -- which factors make an Airbnb popular, which factors make the price change over time, and based on that, we will then forecast the price in the future.

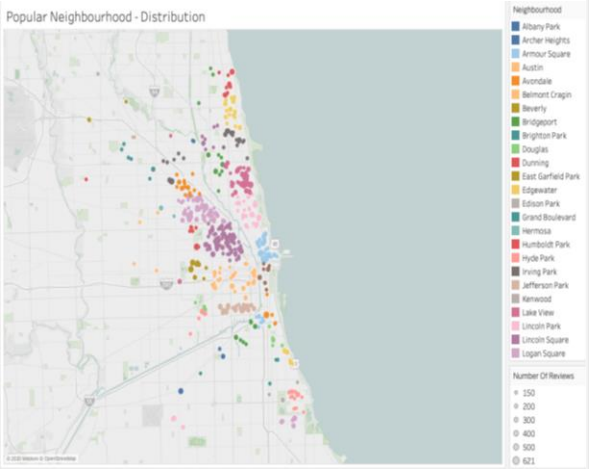
Data Visualization and Objectives

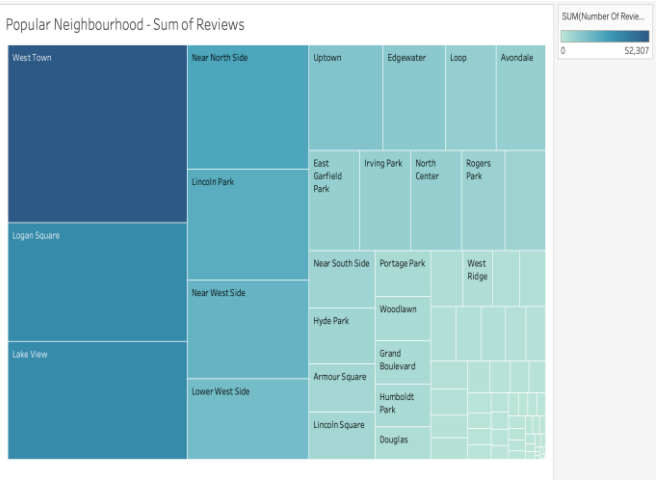
Part 1 – Popularity Analysis

Question 1 – Which neighborhoods are popular?

For the popularity analysis, our assumption is that the higher number of reviews, the more popular an Airbnb is. We first filter popular housing by number of reviews, which should be larger than 150, and then group these housing by different neighborhoods and show them on the map *[Figure-1]*. In this figure, different colors stand for different neighborhoods, and a high density of datapoints shows the certain neighborhood is relatively popular. More specifically, we also draw a treemap of the total number of

reviews within a certain neighborhood to show which one is more popular. Both figures show that the West Town, Logan Square, and Lake View are the top 3 popular neighborhoods.



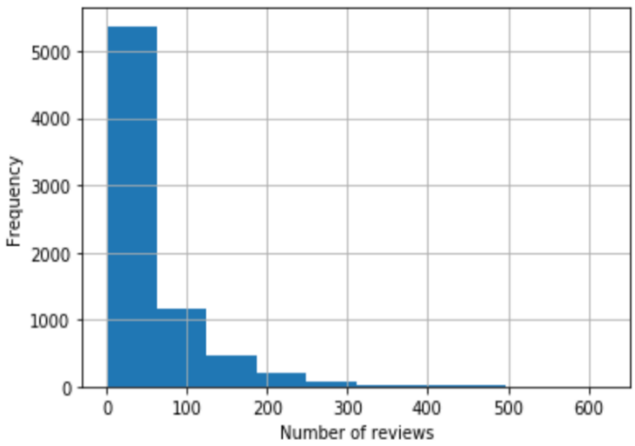


[Figure-1] [Figure-2]

Question 2 - Which factors make an Airbnb or a neighborhood popular?

To better understand the popularity of the housing, we are going to analyze the relationship between the popularity and the attributes of a certain resource, for example, the price, room type, minimum nights, etc. To achieve this goal, we will try some regression tools such as linear model and non-linear model, then we will also try to select the most significant set of predictors among all alternative variables to improve our model.

Logistic regression and model selection



First of all, by plotting the histogram of "number of reviews", we can know the distribution of reviews. Here we label each housing to "popular" and "unpopular" based on the number of reviews (more than or equal to 50 reviews means "popular", and less than 50 reviews means "unpopular"), then we use logistic regression to fit a binary classification model. According to the original dataset, there are plenty of features

we can use as our predictors, but it is unreasonable to use all features as predictors. Hence, in order to improve the interpretation of our model and find out which factors are most significant to popularity, we also try several methods to do model selection.

We first fit a logistic regression model based on all alternative features. There are 5 categorical features (*"zipcode", "host_is_superhost", "property_type", "room_type", "cancellation_policy"),* all of them are treated as dummy variables. It is worth mentioning that we include "zipcode" as one of our predictors. Although, as a categorical predictor, it contains many levels and makes the model more complicated, we think it can capture the impact of location, which is essential to the popularity of a housing. Besides, there are also 6 numerical features we can use (*"accommodates", "bathrooms", "bedrooms", "beds", "price" and "review_scores_rating"*). Putting all of these features into our logistic regression model, we get a model with 71.35% training error and 71.22% test error.

Then we are going to simplify the model and find out what features really matter to the popularity of housing. Hence, we do model selection by LassoCV. By adjusting the threshold, we can control the number of features selected. First, we set a high threshold to ensure only one feature is selected, it turns out the most important feature to popularity is "*host_is_superhost_t*". It means if a host is a super host, then this Airbnb is very likely to be popular. Similarly, after lowering the threshold a little bit, it comes the second important feature, which is "*cancellation_policy_flexible*". It means if the cancellation policy is flexible, then this Airbnb is likely to be popular. Keeping lowering the threshold, we finally find out the top 3 features that make an Airbnb popular, they are "super-host or not", "cancellation policy" and "accommodates".

To verify this result, we refit the logistic regression model based on these three features only. It turns out the classification accuracy is 67.1%, which doesn't reduce too much compared with the model suing all features. Hence, we can safely draw the conclusion that these three features are the most important features to the popularity of housing.

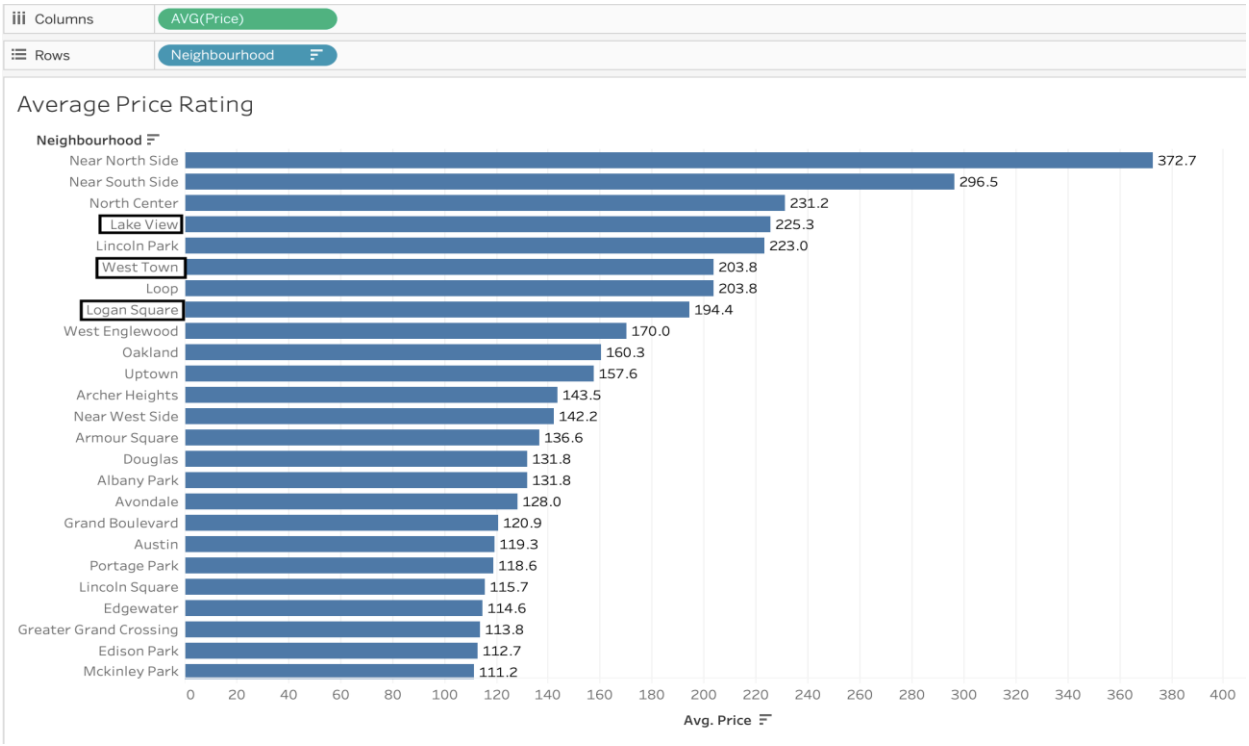
Inspiration

The result shows the top 3 features that are important to popularity do not include "zip code", which we thought would capture the information of location. However, based on domain knowledge, location is a significant feature to the popularity of housing. This indicates that the feature "zip code" might not be precise enough to capture the location information and we probably need to analyze the location information via a non-model-based way.

Part 2 – Price Analysis

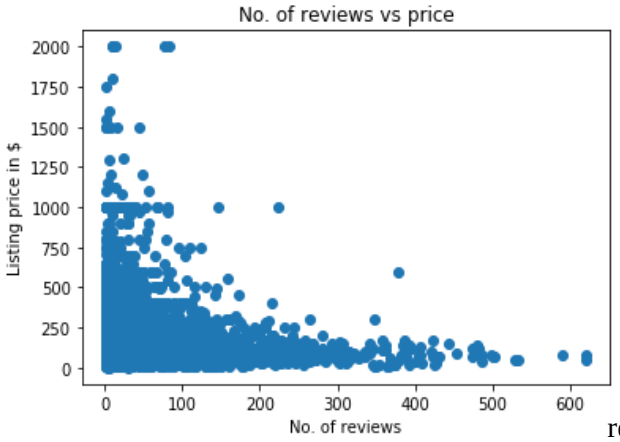
Question1 - Which neighborhood is expensive? How expensive are the most popular neighborhoods?

First, we plot a horizontal bar chart to show the average price of housing for each neighborhood. From this figure, we can also see where the prices of top 3 popular neighborhoods are *[Figure-3]*.



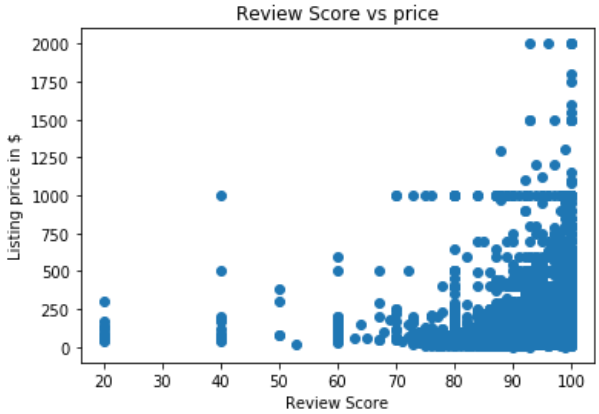
[Figure-3]

Question2 – Is there a relationship between price and popularity?



First, we look at certain properties such as the number of the reviews of each place, and the score that each review left to see their correlation with price. We then create scatterplots of price vs some features, trying to have a sense of what factors might affect the price. From the plot of "Listing price" vs "Number of reviews", we can see an apparent negative correlation in these two variables: the residences of high number of reviews always have low

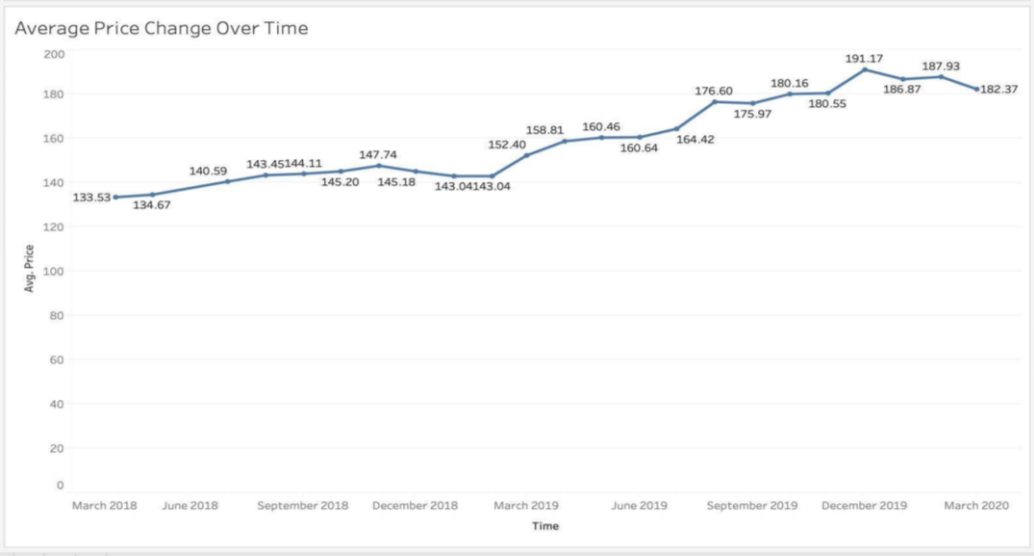
prices, while the expensive residences always have a low number of reviews. This could possibly be explained by the fact that as more people stay at a place, they will leave more reviews, and cheaper places are more accessible to more people. Hence, we hypothesize the number of reviews might be an important factor that affects the price.



Similarly, we can see the plot of "price vs review score". There is a clear positive correlation present, with only a couple outliers. *[Review score vs price]* Clearly, we can explain this as no one would stay at an expensive place with bad reviews, but some cheap places might warrant bad reviews, and both expensive and cheap places could be given nice high reviews.

Question3 – What is the average price of housing in Chicago? How does the price change over time?

To answer the chronological question, we have to explore the archived data for the past few years. Limited by the data we can access, here we collect the archived data of each month from April, 2018 to March, 2020 (data of June, 2018 is missing). For each dataset, we calculate the average price of all housing at that month and then plot a line chart of price change *[Figure-4]*.



[Figure-4]

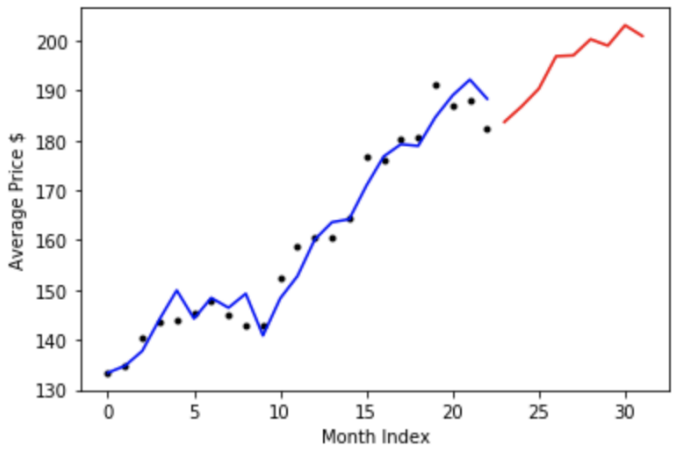
As the line chart shows, the average price of housing has been steadily rising, along with some fluctuations from the last two years. Although the data is not enough, we can still observe some patterns of the price change. For example, the price tends to increase at the end of a year and decrease at the beginning of a year.

Question4 - How will the price change in the future?

Forecasting

Knowing how the price changes over time, we are more interested in the reasons why price change happens and why there is a certain pattern of changing. Therefore, we are going to explore further questions about the price changing and try to forecast the change in the future.

Through the price change of past two years, we find that there is an additive trend and also an additive seasonality among the price change. Hence, we choose to use Holt-Winter method to do forecasting, setting both trend and seasonality to "additive", and predict the average price of Airbnb housing in Chicago in the next 9 months (the rest of the year). The predicted average prices are shown in



Predicted Average Price	
Month	**Ave Price $**
2020/4	183.67
2020/5	186.79
2020/6	190.34
2020/7	196.85
2020/8	197.00
2020/9	200.24
2020/10	198.96
2020/11	203.05
2020/12	200.89

[Figure-4]

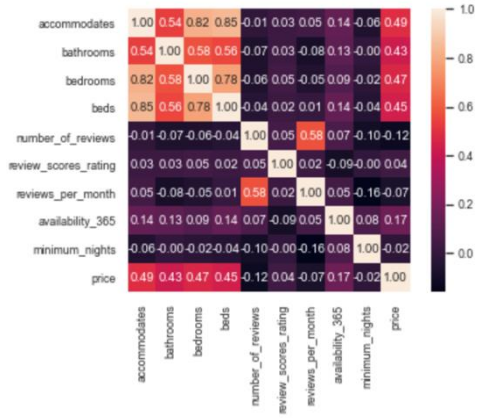
Inspiration

Since we only have the data of the past two years, and we are trying to forecast the price for next couple months based on these data, the result might be not very accurate and reliable, so we are going to go further into the price analysis and trying to find out what factors really matter to the price of Airbnb housing.

Question5 - What factors are most related to the price?

Qualitative analysis

In general, we think the price of housing might be affected by factors such as the location, the ratings, room type, etc. To have a general sense of our alternative features, we plot a correlation matrix to see the correlation between price and other features. Since the correlation matrix can only show correlations among numerical variables, here we only plot the correlation matrix of all alternative continuous variables.



As a result, the correlation matrix does not show strong correlations between price and any numerical features, which indicates it is impossible to get a perfect regression model only based on one numerical feature. However, they are still likely to jointly affect the price. Hence, we decide to explore more features, including categorical features, such as “zip code”, “room type” and “property type”, then do regression analysis and model selection with numerical variables.

Linear regression and model selection

To better understand how Airbnb housing is priced, we then look at all possible factors that might affect price. Instead of the average price, we focus on the price of each residence, and try to use regression and model selection methods to better analyze the price. Before that, we have to modify some parts of our data. Factors such as room type and zip code (capture the information of location) are not originally numerical values, for these we create dummy variables to be able to use linear regression on them.

Firstly, we simply use all alternative features to run a linear regression model. The result shows the R square of the full model is quite low (0.41), which indicates there are some features we should exclude to get a better model. Hence, we then do model selection using 2 model selection methods: Optimizing by P-value and using LassoCV, trying to simplify our model and get a better degree of fitting We then run 2 different types of model selection, trying to find the most important factors that affect the price.

We first look at optimizing our model by filtering through p values. This gives us an $R^2 = 0.664$, and an Adj. $R^2 = 0.663$. By setting p-value threshold to 0.02, it gives out 11predictors with relatively low p-values. They are “accommodates”, “bathrooms”, “bedrooms”, “beds”, “number_of_reviews”,

“reviews_per_month”, “availability_365”, “zipcode_60611”, “property_type_Boutique hotel”, “property_type_Hostel”, “property_type_Hotel”.

To some extent, it can indicate what features are significant in the regression model. However, using p-value to select predictors among a large number of features is not always reliable, so we also have to try another model selection method.

Then, we try using the LassoCV method and changing around the threshold values to find the best model. By setting a threshold equal to 3, it gives out the top 5 features that are most important to the price. They are “accommodates”, “bedrooms”, “bathrooms”, “zipcode_60611”, and “room_type_Private room”. Then, using the selected predictors, we run a new regression model. The result shows the degree of fitting is better than the full model, which indicates the model using 5 predictors is better than the model using all 129 alternative predictors. That verifies that the majority of alternative features have no relationships to the price and only some of them really matter.

From two model selection models, we can conclude that the most important factors in determining the price of a listing are the location, the amenities (bathrooms and bedrooms), and property type.

OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.639
Model:	OLS	**Adj. R-squared (uncentered):**	0.638
Method:	Least Squares	**F-statistic:**	1296.
Date:	Mon, 11 May 2020	**Prob (F-statistic):**	0.00
Time:	15:15:56	**Log-Likelihood:**	-23242.
No. Observations:	3669	**AIC:**	4.649e+04
Df Residuals:	3664	**BIC:**	4.652e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
accommodates	10.7094	1.338	8.003	0.000	8.086	13.333
bathrooms	57.3982	3.603	15.932	0.000	50.335	64.462
bedrooms	20.0661	3.702	5.420	0.000	12.807	27.325
zipcode_60611	119.3112	13.323	8.955	0.000	93.190	145.433
room_type_Private room	-44.2881	4.892	-9.054	0.000	-53.879	-34.698

Omnibus:	3382.344	**Durbin-Watson:**	1.929
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	171052.163
Skew:	4.307	**Prob(JB):**	0.00
Kurtosis:	35.322	**Cond. No.**	35.2

Inspiration

The relatively low value of R square of the linear regression model indicates the fitting of linear model is not perfect in price analysis, which makes the model selection also not reliable enough. However, it can still explain what factors are important to the price of Airbnb housing. To further explore the price, more complicated model, such as non-linear models, should be attempted.

Summary

In this project, we explore the Airbnb listings in Chicago, with focus on two aspects of housing: popularity and price. In both parts, we show some concerned and interesting facts via data visualizations and dig into some questions using different data analysis techniques and better understand the factors that affect the popularity and price of Airbnb housing. We also get some meaningful inspirations after analyses, which can enlighten us and other researchers when working on similar research.