University of Toronto

Analysis of factors affecting 2019 New York City Airbnb listing price

Cong Liu

ECO225: Data Tools for Economists

100626720

April 22, 2021

1.0 Introduction

This paper aims to investigate the key influencing factors behind the 2019 Airbnb rental pricing in New York City, US. In recent years, sharing economy has been rapidly growing. A plethora of digital platforms emerged, enabling asset owners to capitalize the unused capacity of things they already have, and consumers to rent from their peers rather than rent or buy from a company (Geron, 2013). Airbnb is one of the most popular and successful house rental platforms that emerged from the sharing economy. Airbnb describes itself as being able to "help make sharing easy, enjoyable and safe" (Airbnb, n.d.). The digital platform focuses on connecting hosts who have a vacant room and travellers who seek a short-term stay and providing its users a variety of accommodation information. An empirical analysis that included forty major US cities contends that Airbnb listings are more likely to locate in neighbourhoods that are in proximity of the city center, have convenient transportation, and are of high median housing price and household income (Jiao and Bai, 2020). Results from a socio-economic analysis of Airbnb in New York City also highlight that Airbnb houses concentrate on the parts of the city with a large number of housing units and tourist attractions (Dudás, Vida, Kovalcsik and Boros, 2017). Previous researches focused more on the attributes that affect the spatial distribution of Airbnb listings across the city. There are not as many pieces of literature pertaining to the factors affecting Airbnb listing price. Amongst the 14 key price determinants presented in past literature (Chattopadhyay, 2019), the dataset I am working on only includes the review metrics of the Airbnb houses.

Incorporating the previous findings on influencing attributes on Airbnb rental prices with my own experience of using the platform as a traveller, I choose to use neighbourhood_group (the borough where the house is located), room type and availability as the three main predictive covariates of my research on 2019 Airbnb pricing in New York City, US. This paper, comparing with other previous literature, focuses only on determinants of Airbnb pricing in New York city in 2019. The scope of the study is smaller, yet more specific.

The research paper first conducts an explanatory analysis on the chosen variables through summary statistics and visualizations. A further and more advanced analysis is performed through the application of OLS model and regression tree. Section 2.3.2 presents a neat decision tree using justified predictors that yields predictions on 2019 New York City Airbnb rental pricing.

2.0 Analysis

2.1 Data

The dataset that I originally work on is New York City Open Data and comes from Kaggle.com (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data). The dataset provides the Airbnb listing activity and metrics in New York City, US for 2019, including homeowners, geographical location (borough, neighbourhood, longitude and latitude), availability of the house and the time and amount of reviews.

As the project progresses, I choose to merge a dataset about 2019(Q4) New York City real estate value (from https://www.propertyshark.com/mason/market-trends/residential/nyc-all) to complement the Airbnb data. This merging decision is inspired by the conclusion from past literature that Airbnb listings are more likely to locate in neighbourhoods with high median housing (Jiao and Bai, 2020).

2.2 Summary Statistics and Visualization

2.2.1 Variable analysis on Airbnb rental price

Table 1: Summary Statistics of Airbnb rental price

count	31354
mean	162.091822
std	254.444750
min	10
25%	70
50%	112
75%	189
max	10000

There is a huge difference between the 75th quantile of the Airbnb rental price (189 US Dollars) and the maximum of prices (10000 US Dollars). Besides, the average rental price (162.09 US Dollars) is larger than the median price (112 US Dollars). Thus I infer that the distribution of Airbnb rental prices is highly skewed to the right.

To better visualize the distribution of Airbnb rental prices, I calculated the IQR of price and use that as the standard to drop outliers. In align with my previous observation, the histogram (see below) shows a unimodal and right-skewed distribution. The highest frequency values lie approximately between 50 US dollars to 100 US dollars. The price of top frequency is 150 US dollars which is slightly lower than the average price of 162 US dollars.

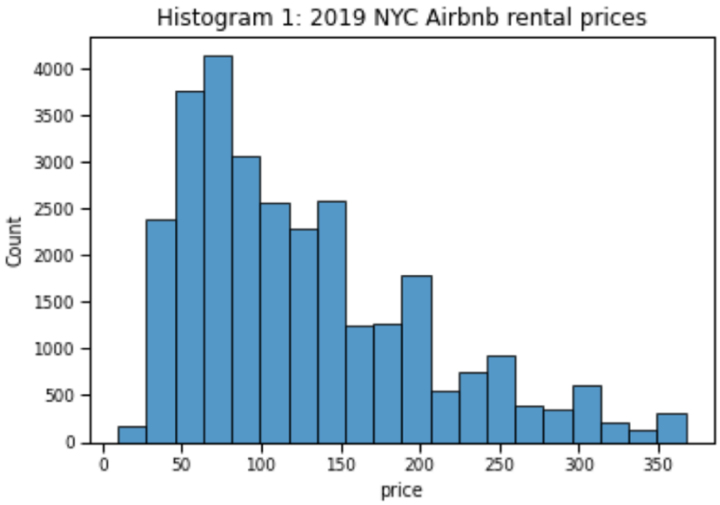

Figure 1: Histogram of 2019 New York City Airbnb rental prices

2.2.2 Variable analysis on Airbnb listing borough

Table 2: Summary Statistics of Airbnb listing borough

count	31342
unique	5
top	Manhattan
freq	13555

Amongst the five boroughs in New York City, Manhattan is the most popular location. The majority of 2019 NYC Airbnb listings (about 43.25%) is located in Manhattan.

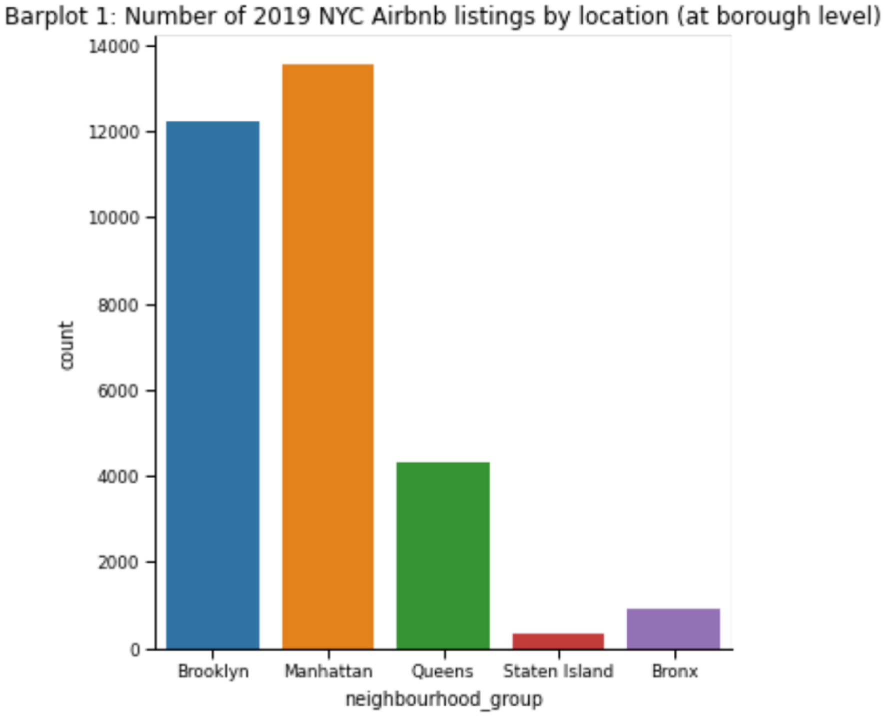

Figure 2: Bar plot of 2019 New York City Airbnb listing locations (at borough level)

In correspond to the summary statistics for Airbnb locations, the NYC borough with the most Airbnb listings is Manhattan, followed by Brooklyn (just over 12000 listings). Staten Island has the fewest (331) Airbnb listings.

2.2.3 Variable analysis on Airbnb listing space type

Table 3: Summary Statistics of Airbnb listing space type

count	31354
unique	3
top	Entire home/apt
freq	16532

The three room types available in the New York City Airbnb are entire home, private room and shared room. The majority (about 52.7%) of houses on Airbnb is available as an entire home or apartment.

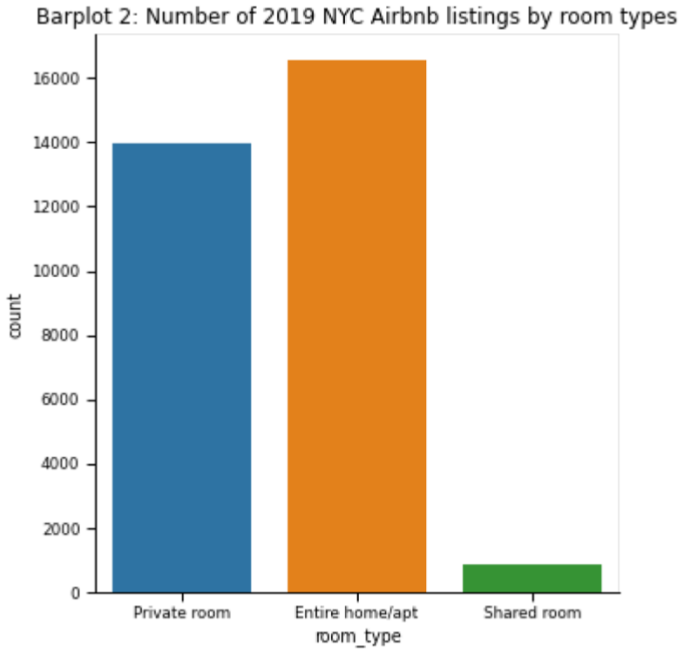

Figure 3: Bar plot of 2019 New York City Airbnb listing space types

Figure 3 shows the same fact that most NYC Airbnb listings are available as an entire home or apartment. Fewer than 1000 houses on New York City Airbnb are available for rental as a shared room.

2.2.4 Variable analysis on Airbnb listing availability

Table 4: Summary Statistics of Airbnb listing availability

count	31342
mean	175.799949
std	126.183544
min	1
25%	55
50%	167
75%	305
max	365

The number of days when the house is available for booking on Airbnb ranges from only one day to a full year. The difference between the 75th quantile and the 25th quantile of housing availability is around eight months (250 days), which indicates a large variability in Airbnb availability. In addition, the average availability (approximately 176 days) is larger than the median (167 days). Hence I presume that the distribution of Airbnb availability is skewed to the right.

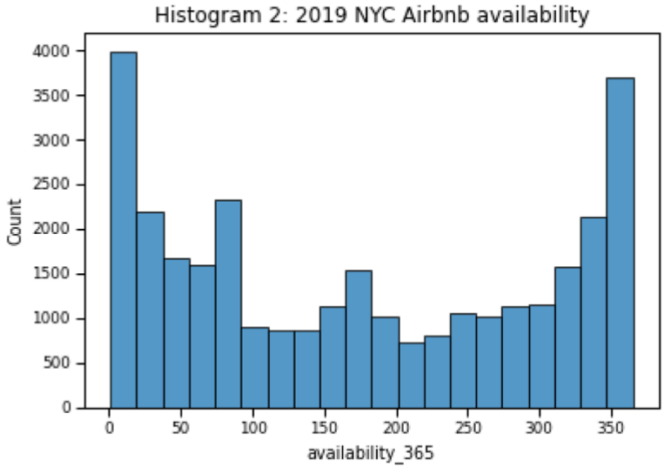

Figure 4: Histogram of 2019 New York City Airbnb listing availability

The histogram of Airbnb availability shows a bimodal and roughly symmetric distribution. The highest frequency values lie approximately at one day and 365 days. The availabilities of top three frequency are 365 days, 364 days and one day, respectively. The most frequent availabilities are divergent from the average NYC Airbnb availability of 176 days.

2.3 Regression Results

2.3.1 OLS Models

Amongst the four regressions using OLS model run in the project, the fitted linear model for the relationship between Airbnb rental price and 2019 (Q4) NYC median housing price has the largest R squared value (regression summary see below). The R squared value of 0.032 indicates that around 3.2% of variation in 2019 New York City Airbnb rental price is explained by the median house price in the borough where the listing is located.

Table 4: OLS Regression Results for the linear model on Airbnb rental price versus median housing price

<table>
<tr><td colspan="7">OLS Regression Results</td></tr>
<tr><td>Dependent Variable:</td><td>price</td><td colspan="4">R-squared:</td><td>0.032</td></tr>
<tr><td>Model:</td><td>OLS</td><td colspan="4">Adjusted R-squared:</td><td>0.032</td></tr>
<tr><td>Method:</td><td>Least Squares</td><td colspan="4">F-statistic</td><td>1045</td></tr>
<tr><td># of observations:</td><td>31342</td><td colspan="4">Prob (F-statistic)</td><td>1.29e-225</td></tr>
<tr><td></td><td>coef</td><td>std err</td><td>t</td><td>P>|t|</td><td colspan="2">[0.025, 0.975]</td></tr>
<tr><td>const</td><td>105.9722</td><td>2.239</td><td>47.327</td><td>0.000</td><td>101.583</td><td>110.361</td></tr>
<tr><td>median_sale_price</td><td>3.136e-05</td><td>9.7e-07</td><td>32.332</td><td>0.000</td><td>2.95e-05</td><td>3.33e-05</td></tr>
</table>

The intercept $\widehat{\beta_0}$=105.97 and the slope $\widehat{\beta_1}$=3.316×10^{-5}. The positive $\widehat{\beta_0}$ parameter estimate implies that the median housing price has a positive effect on Airbnb rental prices. The p-value of 0.00 for $\widehat{\beta_1}$ implies that the effect of median house price on Airbnb rental price is statistically significant using $p < 0.05$ as a rejection rule. Using the parameter estimates, the estimated relationship between Airbnb rental price and 2019(Q4) NYC median housing price is

$$\widehat{\text{price\_i}}=105.67+3.316*10^{-5}\ \text{median\_sale\_price}_i$$

The other three linear regression results are presented below.

Table 5: OLS Regression Results for the linear model on Airbnb rental price versus house availability

OLS Regression Results						
Dependent Variable:	price	R-squared:				0.006
Model:	OLS	Adjusted R-squared:				0.006
Method:	Least Squares	F-statistic				175.0
# of observations:	31342	Prob (F-statistic)				7.73e-40
	coef	std err	t	P>|t|	[0.025, 0.975]	
const	135.6853	2.458	55.193	0.000	130.867	140.504
median_sale_price	0.1503	0.011	13.228	0.000	0.128	0.173

The intercept $\widehat{\beta_0}$=135.69 and the slope $\widehat{\beta_1}$=0.1503. The positive $\widehat{\beta_0}$ parameter estimate implies that the number of days when the house is available for booking has a positive effect on Airbnb rental prices. The p-value of 0.000 for $\widehat{\beta_1}$ implies that the effect of housing availability on Airbnb rental price is statistically significant using $p < 0.05$ as a rejection rule. Using the parameter estimates, the estimated relationship between Airbnb rental price and availability is

$$\widehat{\text{price\_i}}=135.69+0.15*\ \text{availability\_365}_i$$

Table 6: OLS Regression Results for the linear model on Airbnb rental price versus amount of nights minimum

<table>
<tr><td colspan="7">OLS Regression Results</td></tr>
<tr><td>Dependent Variable:</td><td>price</td><td colspan="4">R-squared:</td><td>0.002</td></tr>
<tr><td>Model:</td><td>OLS</td><td colspan="4">Adjusted R-squared:</td><td>0.002</td></tr>
<tr><td>Method:</td><td>Least Squares</td><td colspan="4">F-statistic</td><td>48.51</td></tr>
<tr><td># of observations:</td><td>31342</td><td colspan="4">Prob (F-statistic)</td><td>3.35e-12</td></tr>
<tr><td></td><td>coef</td><td>std err</td><td>t</td><td>P>|t|</td><td colspan="2">[0.025, 0.975]</td></tr>
<tr><td>const</td><td>158.5181</td><td>1.526</td><td>103.886</td><td>0.000</td><td>155.527</td><td>161.509</td></tr>
<tr><td>median_sale_price</td><td>0.4317</td><td>0.062</td><td>6.965</td><td>0.000</td><td>0.310</td><td>0.553</td></tr>
</table>

The intercept $\hat{\beta}_0$=158.52 and the slope $\hat{\beta}_1$=0.43. The positive $\widehat{\beta_0}$ parameter estimate implies that the amount of nights minimum has a positive effect on Airbnb rental prices. The p-value of 0.000 for $\widehat{\beta_1}$ implies that the effect of minimum nights on Airbnb rental price is statistically significant using $p < 0.05$ as a rejection rule. Using the parameter estimates, the estimated relationship between Airbnb rental price and minimum nights is

$$\widehat{price\_i}=158.52+0.43*\ minimum\_nights_i$$

Table 7: OLS Regression Results for the linear model on Airbnb rental price versus number of reviews per month

OLS Regression Results						
Dependent Variable:	price	R-squared:				0.007
Model:	OLS	Adjusted R-squared:				0.007
Method:	Least Squares	F-statistic				215.3
# of observations:	31342	Prob (F-statistic)				1.41e-48
	coef	std err	t	P>|t|	[0.025, 0.975]	
const	179.9177	1.878	95.811	0.000	176.237	183.590
median_sale_price	-11.7997	0.804	-14.672	0.000	-13.376	-10.223

The intercept $\widehat{\beta_0}$=179.92 and the slope $\widehat{\beta_1}$=-11.80. The positive $\widehat{\beta_0}$ parameter estimate implies that the number of reviews per month has a positive effect on Airbnb rental prices. The p-value of 0.000 for $\widehat{\beta_1}$ implies that the effect of reviews per month on Airbnb rental price is statistically significant using $p < 0.05$ as a rejection rule. Using the parameter estimates, the estimated relationship between Airbnb rental price and number of reviews per month is

$$\widehat{price\_}i=179.92-11.80*\ reviews\_per\_month_i$$

From my personal evaluation, none of the regression results above is satisfactory enough. The adjusted R-squared values for all the models are too small to be explanatory despite the fact that all models have a statistically significant linear association. In addition, both AIC and BIC are on the scale of one hundred thousand for all models, indicating a large prediction error for four models.

To summarize, both the linear relationship of 2019 NYC Airbnb rental price versus minimum nights and the relationship between the rental price and house availability are positive despite weak. The relationship between Airbnb price and the number of reviews is interestingly negative, which verifies my previous theory that Airbnb users tend to leave reviews for non-satisfying stays. In response, hosts lower the rental price to attract more travellers. The strongest positive linear relationship between Airbnb rental price and the median house sale price adds onto the conclusion from past literature. Airbnb listings located in a neighbourhood with a high median housing price are very likely to have a high rental price point.

2.3.2 Regression Tree

For the regression tree presented below, my objective function is MSE (Mean Square Error), which is the most commonly used measurement of accuracy:

$$\frac{1}{N}\sum_{i=1}^{N}\left((\text{price}_i)\text{-}\left(\beta_0+\beta_1\text{*median\_sale\_price}_i\right)\right)^2$$

My goal is to minimize the average squared deviation between the actual Airbnb rental price and the predicted price from the OLS model in 2.3.1 that uses median housing price as the predictor.

An existing problem in the chosen linear model is multicollinearity despite its relatively large R-squared value. Under this scenario, a regression tree seems to be more appropriate to make predictions. A trustworthy regression tree is expected to be highly accurate while not too complex. To resolve overfitting and/or multicollinearity of a linear regression model, a possible solution is to "penalize", that is, to reduce the magnitude or values of the relatively insignificant variables. Regularization is able to kill two birds with one stone. The parameter responsible for cutting off complexity and improving model quality is known as the regularization parameter. A larger tuning parameter means a larger penalty for having a complex tree. Thus $\alpha = \infty$ means tree pruning at the highest strength, yielding a subtree of optimal size.

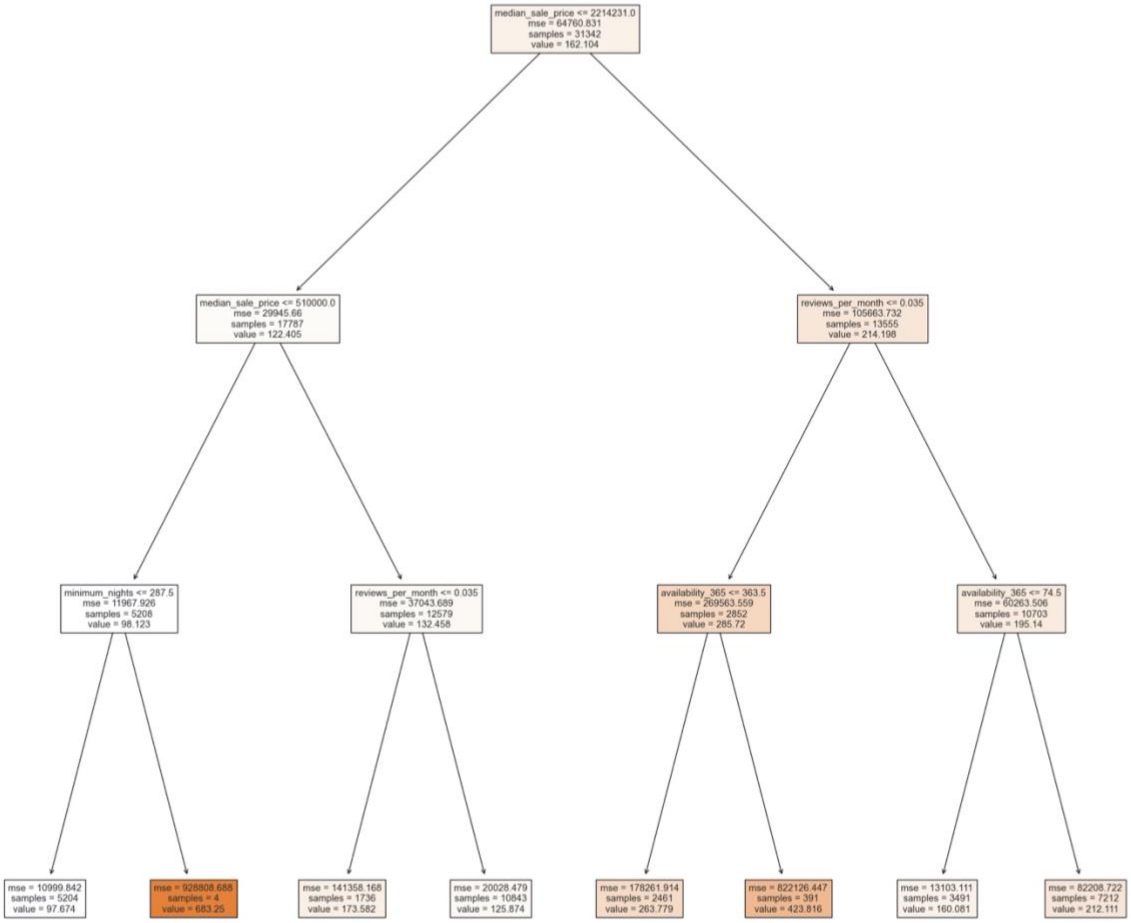

Figure 5: Regression Tree for 2019 NYC Airbnb rental price prediction

The above is a regression tree for predicting 2019 NYC Airbnb rental prices, based on the Airbnb listing availability, amount of nights minimum, number of review per month and the median housing price. The tree has six internal nodes and eight terminal nodes. Each internal node includes the predictor used, MSE associated with this decision, number of observations that lands in this node and the predicted rental price value. Each terminal node includes the same information and excludes the involved predictor.

MSE of the regression tree with depth of 3 (61333.395035149726) is better than simple linear regression including all four Xs (62078.90567522036). This indicates that prediction yielded by the regression tree is less error-prone comparing with a linear regression.

At the root, the predicted price from the chosen OLS model (see Table 4) seems to overshoot (175.4 US dollars versus 162.104 US dollars from the tree). The overestimate of the linear model could be due to the lower number of features included. Yet the rental price predictions yielded by two models are quite close at the internal node of the regression tree (*median_sale_price <= 510000*).

3.0 Conclusion and Future Work

In Section 2.2, I briefly explored how location, room type and availability (i.e. `neighbourhood_group`, `room_type` and `availability_365`) affect the 2019 NYC Airbnb rental prices.

Looking at the variables individually, we can see that the distribution of prices of NYC Airbnb listings is highly skewed. The majority of Airbnb listings in New York city is located in Manhattan. Also, most of the houses on NYC Airbnb are rented out as an entire home. On the other hand, the distribution of availability of Airbnb listings is bimodal. Most houses are available for only one day or almost a whole year.

The scatterplots show a very weak, positive linear relationship between price and `availability_365`. As the house is available for a longer time, the rental price rises on average. This observation is in align with my intuition.

However, the correlation coefficient indicates that there is a weak negative association between price and longitude. That is, as longitude increases, price is expected to fall on average. Whereas an increase in longitude implies going further in the East direction. In the future, I would want to analyze why prices of Airbnb listings in New York City show this trend. Does this trend hold in other cities or regions?

Amongst the boxplots of Airbnb rental price across the five New York city boroughs, the one for Manhattan is highly similar with the overall price distribution. It is very likely that the similarity originates from the fact that most Airbnb listings are located in Manhattan. Rental prices of Airbnb houses in Manhattan also show the largest variability.

In the future, review metrics shall also be included into consideration as they appear to be linearly associated with price. I could do a time-series analysis on the time when reviews of

Airbnb listings are posted. Using the web-scraping technique, I could conduct a text analysis, extract sentiment from the comments and study its relationship with Airbnb rental price.

References

Airbnb, Inc. “What is Airbnb and how does it work?”. Accessed April 18, 2021. https://www.airbnb.com/help/article/2503/what-is-airbnb-and-how-does-it-work

Chattopadhyay, M., & Mitra, S. K. (2019). Do airbnb host listing attributes influence room pricing homogenously?. *International Journal of Hospitality Management, 81*, 54-64.

Data.gov. "Borough Boundaries (Water Areas Included)". Accessed February 23, 2021. https://catalog.data.gov/dataset/borough-boundarieswater-areas-included

Dudás, G., Vida, G., Kovalcsik, T., & Boros, L. (2017). A socio-economic analysis of Airbnb in New York City. *Regional Statistics*, *7*(1), 135-151.

Geron, Tomio. “Airbnb And The Unstoppable Rise Of The Share Economy”. Accessed April 18, 2021. https://www.forbes.com/sites/tomiogeron/2013/01/23/airbnb-and-the-unstoppable-rise-of-the-share-economy/?sh=5a526021aae3

Jiao, J., & Bai, S. (2020). An empirical analysis of Airbnb listings in forty American cities. *Cities*, *99*, 102618.

Kaggle. "New York City Airbnb Open Data". Accessed January 31, 2021. https://www.kaggle.com/dgomonov/new-york-city-airbnb-opendata

Kaggle. "Lin Regression Hospitality in Era of Airbnb Code_3". Accessed March 22, 2021. https://www.kaggle.com/biphili/lin-regressionhospitality-in-era-of-airbnb-code-3

pandas development team. "API reference". Accessed January 31, 2021. https://pandas.pydata.org/pandas-docs/stable/reference/index.html

PropertyShark.com. “NYC Real Estate Market Trends”. Accessed April 18, 2021. https://www.propertyshark.com/mason/market-trends/residential/nyc-all

StackOverFlow. "how to use pandas filter with IQR?". Accessed March 2, 2021.

https://stackoverflow.com/questions/34782063/how-touse-pandas-filter-with-iqr

TED. "The economy of trust". Accessed February 1, 2021. https://www.ted.com/playlists/366/the_economy_of_trust

U.S. Department of Commerce. “QuickFacts”. Accessed February 23, 2021. https://www.census.gov/quickfacts/fact/table/US#

Waskom, Michael. "API reference". Accessed February 1, 2021. http://seaborn.pydata.org/api.html

Wikipedia. “Airbnb.” Last modified April 17, 2017. http://en.wikipedia.org/wiki/Airbnb

Wikipedia. “Decision tree pruning”. Last modified April 20, 2021. https://en.wikipedia.org/wiki/Pruning_%28decision_trees%29