Potential Gentrification Trends of New York City Airbnbs and Listing Price Prediction Using Machine Learning

Josh Kim

1 Introduction

Airbnb is one of the world's largest marketplaces for house rentals and activities to do, offering over 7 million accommodations and 50,000 handcrafted activities. They are all powered by local hosts and are accessible in 62 languages across more than 220 countries and regions.[1] In the past 10 years, Airbnb has received criticism that it is triggering gentrification in major cities and driving up short-term rental prices. In particular, this paper focuses on Airbnb listings in New York City because it has the most population in the U.S. and also it is the main city that students from my college, Vassar College, visit for career and entertainment. There exist abundant literature which focused on various socioeconomic impact Airbnbs had on New York City such as gentrification until 2017. But not many papers examined how the gentrification scene looks in New York City since 2017. This paper aimed to fill that gap by analyzing data from 2016-2020 and pointing to areas with a high potential of gentrification. Moreover, I created statistical models predicting New York City Airbnb listing to beat predictive performance from previous research and to provide insights on which factors play a pivotal role in high or low listing prices.

2 Previous Literature and Projects

Previous literature looked into various socioeconomic implications of Airbnbs including the relationship between ethnicity of hosts and revenue, impact on short-term rentals, geographical distribution of listing concentrations and how they contribute to gentrification in major cities.

In regards to listing price prediction, I identified two previous researches– a paper named *Airbnb Price Prediction Using Machine Learning and Sentiment Analysis* and a personal project from a current data scientist at Uber. That paper touches upon a multitude of methodologies including feature selection, regularized linear regression combined with K-Means clustering, Gradient Boosting Tree Ensemble, and Neural Networks with the best model having R Squared score of 69% and MSE defined on ln(price) of 0.147. However, the dataset used is more extensive with more than 90 features unlike my dataset less than 30 basic features.[2] Next, the personal project from Samuel Lam, again, used a similar dataset with more than 50 features and the best model scored $21.43 for its Median Absolute Error.[3] I aim to beat these best scores these previous research attained. The regression metric I tried to minimize was the Median Absolute Error which does not give heavy penalty to large errors. This is because I argue that the best model in this context should predict well the prices of overall listings instead of giving more weight to some outlier listings. Nevertheless, I recorded three other metrics – Mean Absolute Error, R Squared, and Mean Squared Error.

3 Data

The main source of data comes from Inside Airbnb, an independent, non-commercial set of tools and data that allows users to explore how Airbnb is really being used in cities around the world.[4] According to the website, it is not associated with or endorsed by Airbnb or any of Airbnb's competitors. I collected 5 data sets from the website, each of them from year 2016 to 2020 respectively. In addition, I collected the Neighborhood Tabulation Areas (NTA) data[5]. This data contains geospatial shape information that I needed to draw a map visualization on a more granular level.

4 Methodology

To examine price distributions by room type and boroughs over time, I created side-by-side boxplot visualizations. Furthermore, I ran pairwise OLS regressions of price on room type and boroughs in 2017 and 2020 to see if there was any change in the estimated coefficient and statistical significance of variables. "Pairwise" means, for example, running a regression of price on an indicator variable where 1 stands for entire home/apt listings and 0 stands for private room listings for a certain year. For 2020, in particular, I ran another regression after adding other variables including minimum number of nights required to stay, number of reviews, number of available days, and the number of listings of the host to see how the estimated coefficient change.

[1] About us, Airbnb, https://news.airbnb.com/about-us/.

[2] Airbnb Price Prediction Using Machine Learning and Sentiment Analysis, Kalehbasti, Nikolenko, and Rezaei.

[3] Airbnb Pricing Predictions, Lam, https://airbnb-pricing-prediction.herokuapp.com/.

[4] About Inside Airbnb, Inside Airbnb, http://insideairbnb.com/get-the-data.html.

[5] Neighborhood Tabulation Areas (NTA), https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq.

Table 1: OLS Regressions of price ran on room type and neighborhood groups

Year(s) Considered	Y	X	# of Regressions
2017, 2020	Price	room type (entire house/apt. v.s. private rooms)	2
2020	Price	- room type (entire house/apt. v.s. private rooms)- neighborhood group - minimum number of nights - number of reviews - number of available days - number of listings	1
2017, 2020	Price	room type (private rooms v.s. shared house)	2
2020	Price	- room type (private rooms v.s. shared house) - neighborhood group - minimum number of nights - number of reviews - number of available days - number of listings	1
2017, 2020	Price	neighborhood group (Manhattan v.s. Queens)	2
2020	Price	- neighborhood group (Manhattan v.s. Queens) - minimum number of nights - number of reviews - number of available days - number of listings	1
2017, 2020	Price	neighborhood group (Brooklyn v.s. Queens)	2
2020	Price	- neighborhood group (Brooklyn v.s. Queens) - minimum number of nights - number of reviews - number of available days - number of listings	1
2017, 2020	Price	neighborhood group (Manhattan v.s. Brooklyn)	2
2020	Price	- neighborhood group (Manhattan v.s. Brooklyn) - minimum number of nights - number of reviews - number of available days - number of listings	1

Note: Total of 15 regressions.

Next, I investigated the geographical distributions of luxurious Airbnb listings in 2017 and 2020 to see if there was a change in where those listings were concentrated in. I defined outliers (a.k.a. luxurious listings) as listings whose prices are more expensive than the Q3 + Interquartile Range x 1.5 value. Moreover, I calculated percentages of luxurious listings by boroughs and by neighborhoods. I identified which neighborhoods arose as new areas with high proportions of luxurious listings in 2020 after 3 years have elapsed since 2017.

To observe how the distribution of the number of listings in different areas changed over time, I created visualizations including bar plots by room type over time, horizontal bar plots by boroughs over time, line plots for each borough over time on the same grid, all of which have the y-axis as the number of listings per year. In addition, I drew a map visualization whose individual dot corresponds to each listing and boroughs differentiated by colors. Lastly, I created heat maps for both 2017 and 2020 on a tract level where the color of each tract is darker if it has more listings than others.

Next, I examined which neighborhoods had the highest minimum revenue in 2017 and check if they match with the neighborhoods that experienced gentrification or are in the process of active gentrification. This allowed me to verify if Neil Smith's (1979) rent gap model mentioned in previous literature holds true in real world. The rent gap model describes a situation where the actual economic returns to properties decline while potential economic returns tend to increase. In neighborhoods where this gap between actual and potential returns increases, the result is an increasing incentive for real estate capital to direct new housing investment flows. These investment flows trigger housing prices to blow up and, as a result, attract more affluent newcomers, thereby displacing existing poorer residents.[6] However, since I did not have revenue data for each listing, I

[6] Airbnb and Rent Gap: Gentrification Through the Sharing Economy, Wachsmuth and Weisler.

calculated the minimum revenue with the following formula because previous literature mentioned that the number of user reviews works well as a proxy for Airbnb demand as the completion rate for reviews over the number of stays in an Airbnb accommodation is more than 70% (Quattrone et al. 2016).[7]

Minimum Revenue per day = Price per day x Minimum Number of Nights x Number of Reviews (Demand)

For predictive modelling of listing price, I had features including the 6 features I created from the "name" column (e.g. average word length). I used 2016-2019 data as the training set and the 2020 data as the test set. First, I tested linear models including Simple Linear Regression, Ridge, Lasso and Elastic Net as baseline models with five-fold cross validation on the training set. Next, I tuned hyper-parameters for those linear model baseline models to improve performance. To boost up model performance even further, I performed feature selection by using Lasso's inherent feature selecting characteristic and feature importance functionality of Random Forest. More specifically, two new datasets were created; one dataset was created by removing features with coefficients of zero were removed for Lasso Regression and another dataset was created by removing features with zero scores for Random Forest's feature importance. Then, I tested the linear models and various CART algorithms (e.g. Random Forest, XGBoost, LGBM) on the new feature selected datasets. For feature interpretability, I used two methods – feature importance function of Random Forest and Permutation Importance. Feature importance is determined by how often a certain feature is used to decrease the impurity measure of Random Forest. Permutation importance works a bit differently. It first trains the entire dataset and measures the baseline score. Next, each feature is randomly shuffled and the new data is trained again. The bigger the difference between the score from the unshuffled original data and that from the new data where one feature was randomly shuffled, the more important that feature is.

5 Results and Discussion of Exploratory Data Analysis

5.1 Price Distributions by room type in 2016-2020

The average price for private room listings has been increasing persistently in 2017-2020. In comparison to the average listing price of 2018, the average price of private room listings in 2020 increased by more than 40%. The median price did not change much, staying within a $5 range for all three room type listings. The only room type with increases in both the median and average price in the past year was hotel room. Note that hotel room type started appearing from 2019 data because Airbnb allowed hotels to add left over rooms to its listings since 2019. In addition, regression results show that there was a steeper increase in average price of private rooms than entire home or apartment listings over the past 3 years. In 2017, entire home/apartment listings were, on average, $115.22 more expensive than private room listings with a p-value of less than 2e-16. In 2020, entire home/apartment listings were, on average, $90.01 more expensive than private room listings with a p-value of less than 2e-16. That average price difference between entire home/apartment and private room listings was even smaller ($74.83) when other variables including neighborhood group (e.g. Manhattan, Brooklyn), minimum number of nights required to stay, number of reviews, number of available days, and the number of listings of the host were taken into account.

On the other hand, the average price difference between private room and shared room listings was consistently minimal in the past 4 years. Private room listings were, on average, more expensive than shared room listings in the $10 - $15 range with a p-value greater than 0.1. That average price difference jumped to $46.70 in 2020, only when other variables including neighborhood group (e.g. Manhattan, Brooklyn), minimum number of nights required to stay, number of reviews, number of available days, and the number of listings of the host were taken into account.

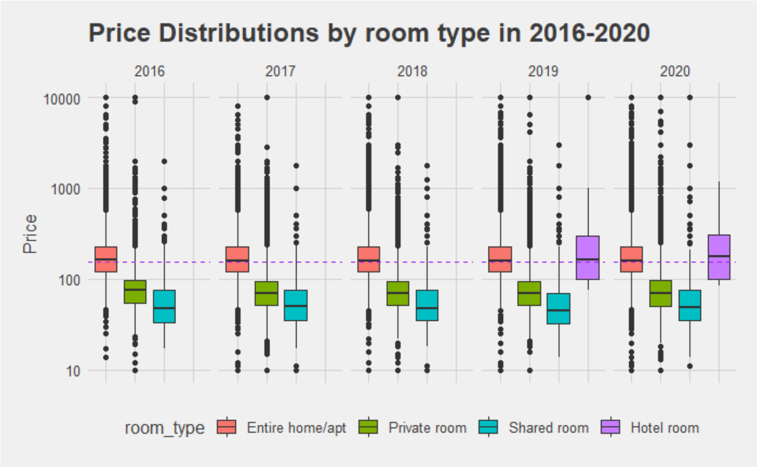
Table 2: Average and Median Prices of Airbnb Listings by Room Type in 2016-2020

		2016	2017	2018	2019	2020
Entire Home/Apt	Average	$206.11	$201.04	$211.35	$210.06	$212.93
	Median	$165	$159	$160	$160	$160
Private Room	Average	$87.19	$85.82	$84.90	$98.26	$122.92
	Median	$75	$70	$70	$70	$70
Shared Room	Average	$67.39	$74.96	$75.33	$75.79	$110.85
	Median	$48	$50	$48	$45	$49
Hotel Room	Average				$271.24	$279.54
	Median				$165	$179

Note: Green indicates not much difference in average prices over time while red indicates considerable changes (more than 40% increase/decrease in 2 years) in average price over time.

[7] A Socio-Economic analysis of Airbnb in New York City, Boros, Dudas, Vida, and Kovalcsik.

Figure 1: Boxplots of Price by Room Type in 2016-2020



Note: y-axis has been log scaled. The purple line indicates average price of all listings in 2016-2020.

5.2 Price Distributions by Neighborhood Group in 2016-2020

The average price of listings was the most expensive in Manhattan followed by Brooklyn with Queens, Bronx and Staten Island tailing on a similar price range. The average price of listings in Manhattan and Bronx increased more than 20% in 2016-2020. Average price of Bronx listings has been steadily increasing and is only $5 less than that of Queens in 2020.

Regression results show that the average price difference between Manhattan and Queens listings has been increasing persistenty over time with those price differences being statistically significant for all years with p-values less than 2e-16. Similarly, the average price difference between listings in Brooklyn and Queens has been increasing consistently over time with those price differences being statistically significant for all years. The average price difference between Manhattan and Brooklyn has been increasing over time with those price differences being statistically significant for all years. This illustrates that the average price of listings in Manhattan has been increasing more steeply than that of listings in Brooklyn.

Table 3: Average and Median Prices of Airbnb Listings by Neighborhood Groups in 2016-2020

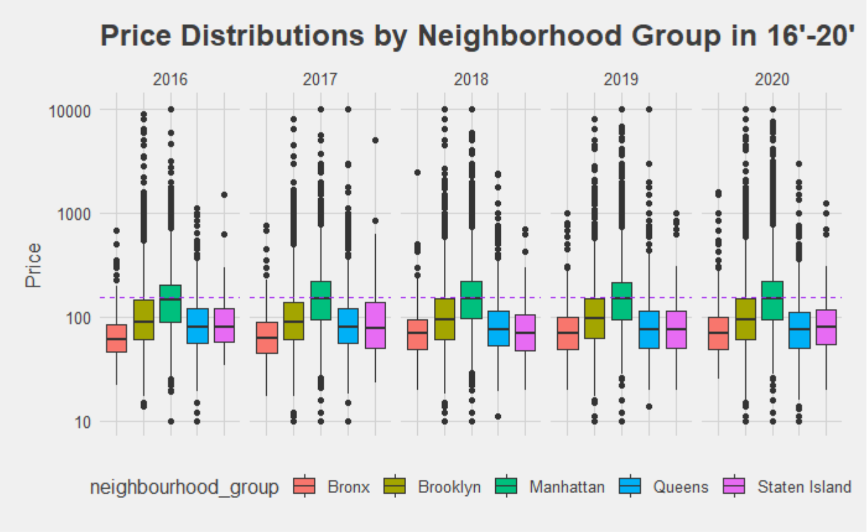
		2016	2017	2018	2019	2020
Manhattan	Average	$183.08	$190.36	$195.17	$206.08	$234.95
	Median	$145	$149	$150	$149	$150
Brooklyn	Average	$121.14	$117.17	$124.43	$127.19	$128.31
	Median	$90	$90	$95	$97	$95
Queens	Average	$99.79	$108.49	$100.54	$105.48	$97.66
	Median	$80	$80	$75	$75	$75
Bronx	Average	$76.07	$78.07	$88.12	$89.77	$92.65
	Median	$60	$63	$70	$69	$69
Staten Island	Average	$142.5	$185.58	$98.04	$108.01	$110.17
	Median	$80	$77.5	$69	$75	$80

Note: Red indicates considerable changes (more than 20% increase/decrease in 4 years) in average price over time.

Table 4: Summary Statistics of Regressions of Price on Neighborhood Groups

Year	Y	X	Variable of Interest	Estimated Coefficient	P-Value
2017	Price	neighborhood group_Queens (base group: Manhattan)	neighborhood Group_Queens (base group: Manhattan)	-81.89	< 2e-16
2019	Price	neighborhood group_Queens (base group: Manhattan)	neighborhood Group_Queens (base group: Manhattan)	-100.59	< 2e-16
2020	Price	neighborhood group_Queens (base group: Manhattan)	neighborhood Group_Queens (base group: Manhattan)	-137.29	< 2e-16
2020	Price	neighborhood group_Queens (base group: Manhattan) minimum number of nights number of review number of available days number of listings	neighborhood Group_Queens (base group: Manhattan)	-148.17	< 2e-16
2017	Price	neighborhood group_Queens (base group: Brooklyn)	neighborhood group_Queens (base group: Brooklyn)	-8.68	< 2e-16
2019	Price	neighborhood group_Queens (base group: Brooklyn)	neighborhood group_Queens (base group: Brooklyn))	-21.71	< 2e-16
2020	Price	neighborhood group_Queens (base group: Brooklyn)	Neighborhood group_Queens (base group: Brooklyn)	-30.65	< 2e-16
2020	Price	neighborhood group_Queens (base group: Brooklyn) minimum number of nights number of review number of available days number of listings	Neighborhood group_Queens (base group: Brooklyn)	-31.80	< 2e-16

Figure 2: Boxplots of Price by Neighborhood Groups in 2016-2020



Note: y-axis has been log scaled. The purple line indicates average price of all listings in 2016-2020.

5.3 Outliers – Luxurious Airbnb Listings

I investigated the geographical distributions of luxurious Airbnb listings in 2017 and 2020 to see if there was a change in where those listings were concentrated in. I defined outliers (a.k.a. luxurious listings) as listings whose prices are more expensive than the Q3 + Interquartile Range x 1.5 value. According to [table 5] below, 66.7% of the luxurious listings were in Manhattan in 2017 and that percentage soared to 73.8% in 2020. Similarly, the percentage of luxurious listings in Bronx increased from 0.4% to 0.6%. These two boroughs were the same two boroughs that recorded increase of more than 20% in average listing price from 2016 to 2020. This indicates that increase in luxurious listings being posted in Airbnb in Manhattan and Bronx was driving up the average price of Airbnbs in those two boroughs.

Table 5: Percentage Breakdown of Luxurious Listings by Neighborhood Groups / Boroughs

	2017	**2020**
Manhattan	66.7%	73.8%
Brooklyn	28.6%	22.2%
Queens	4.1%	3.2%
Bronx	0.4%	0.6%
Staten Island	0.2%	0.2%

Note: Red indicates the percentage of luxurious listings increased from 2017 to 2020 for that borough.

Table 6: Percentage of Luxurious Listings for Top 20 Luxurious Neighborhoods in 2017

Name of Neighborhood	**Percentage(%) of Luxurious Listings**
Midtown	13.4%
Chelsea	11.8%
Williamsburg	9.6%
Upper West Side	7.3%
East Village	6.6%
Hell's Kitchen	5.2%
Lower East Side	4.6%
Financial District	3.9%
Bedford-Stuyvesant	3.9%
Kips Bay	3.8%
Clinton Hill	3.0%
Chinatown	2.3%
Murray Hill	2.2%
Park Slope	2.0%
Flatiron District	1.8%
Bushwick	1.4%
Carroll Gardens	1.4%
Crown Heights	1.3%
Astoria	1.2%
Boerum Hill	1.1%

Table 7: Percentage of Luxurious Listings for Top 20 Luxurious Neighborhoods in 2020

Name of Neighborhood	**Percentage(%) of Luxurious Listings**
Midtown	13.4%
Hell's Kitchen	8.2%
Upper West Side	7.1%
East Village	5.4%
Williamsburg	5.1%
Upper East Side	4.7%
Chelsea	4.6%
West Villiage	3.9%
Harlem	3.0%
Bedford-Stuyvesant	3.0%
SoHo	2.8%
Lower East Side	2.2%
Tribeca	2.0%
Kips Bay	1.9%

Greenpoint	1.8%
Financial District	1.7%
Theater District	1.6%
Crown Heights	1.5%
Murray Hill	1.5%
Nolita	1.5%

Note: Green indicates neighborhoods that were in the top 20 luxurious neighborhoods list for both 2017 and 2020 and also had an increase in the percentage of luxurious listings from 2017 to 2020. Red indicates neighborhoods that were not in the top 20 luxurious neighborhoods list in 2017 but newly appeared in the top 20 luxurious neighborhoods list in 2020.

One previous literature from 2017 alluded to the fact that Airbnb activity overlaps with gentrification as the most popular Airbnb neighborhoods are gentrifying or are already gentrified (e.g. Chelsea, Williamsburg, Greenpoint, Lower East Side, etc.)[8]. Another literature from 2018 explained that neighborhoods including Times Square area, Lower East Side, and Williamsburg were post gentrification areas where landlords have shifted housing supply into short-term rentals to capitalize on the rent gap. It also added that second-tier neighborhoods of Harlem in North Manhattan and Bedford-Stuyvesant in Brooklyn are areas where there were not a lot of Airbnb activity in absolute terms, but where the landlords who are using Airbnb are making a lot more money than they would have in the long-term rental market.[9]

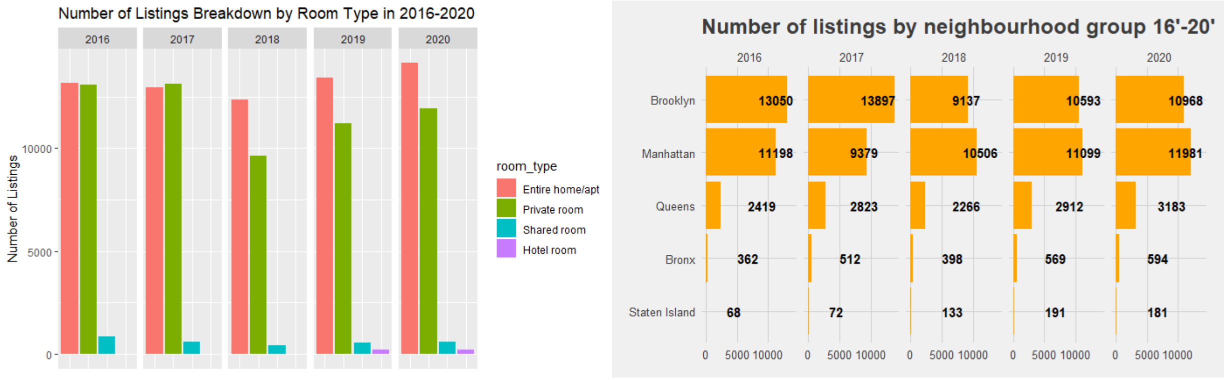
Combining information from [Table 6] and what previous literature have found, I observed that neighborhoods including Midtown, Chelsea, Williamsburg and Lower East Side which were considered to be gentrified already in 2017 had the highest percentage of luxurious Airbnb listings. This suggests potential correlation between having high proportion of luxurious listings and the degree of gentrification that happened in those areas.

In a similar logic, the neighborhoods marked in red in [Table 7], which are areas that newly appeared in the top 20 luxurious neighborhoods list in 2020, may be places where gentrification has been happening in 2017-2020 or already entered the post gentrification stage as of 2020. Those neighborhoods are Upper East Side, West Village, Harlem, SoHo, Tribeca, Greenpoint, Theater District and Nolita. There are two interesting traits about these neighborhoods. First, SoHo, Tribeca and Nolita are all situated in Lower Manhattan which wasn't the most gentrified area in 2018 compared to Midtown. The emergence of these three neighborhoods in Lower Manhattan as luxurious neighborhoods in 2020 may be a sign that gentrification has been intensifying or is almost complete in Lower Manhattan. Second, I wanted to give special attention to Upper East Side and West Village which barely had any luxurious listings in 2017. These two neighborhoods are, as the names suggest, adjacent to Upper West Side and East Village. Upper West Side and East Village were already gentrified to a certain extent and had the highest percentage of luxurious listings combined in 2017. I argue that the emergence of Upper West Side and East Village as new hotspots for luxurious Airbnb listings may demonstrate that once gentrification is complete in one neighborhood, it tends to spill over to the adjacent region.

5.4 Distribution of Number of Listings by Boroughs and Neighborhoods in 2016-2020

The bar plots in figure 3 show that the number of home/apt and private room listings in New York City being posted has been increasing. These increases correspond to the steady surge in number of listings in Manhattan and Brooklyn in 2018-2020 considering majority of the home/apt and private room listings are concentrated in those two boroughs.

Figure 3/4: Number of Listings by Room Type / Number of Listings by Boroughs (Both in 2016-2020)



[8] A Socio-Economic analysis of Airbnb in New York City

[9] Airbnb and Rent Gap: Gentrification Through the Sharing Economy

Figure 5/6: Number of Listings by Boroughs in 2016-2020 (Manhattan, Brooklyn / Queens, Bronx, SI)

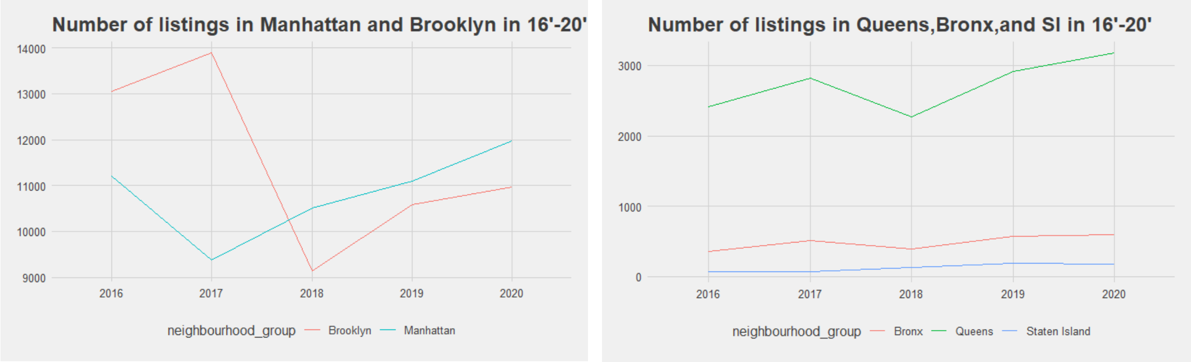
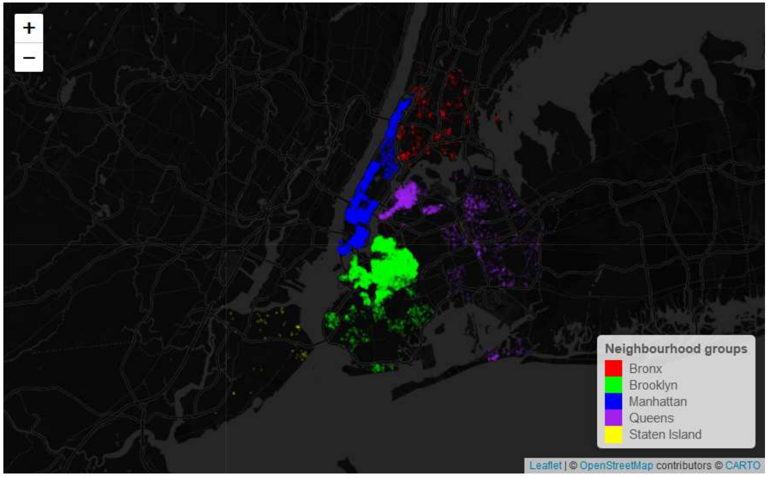


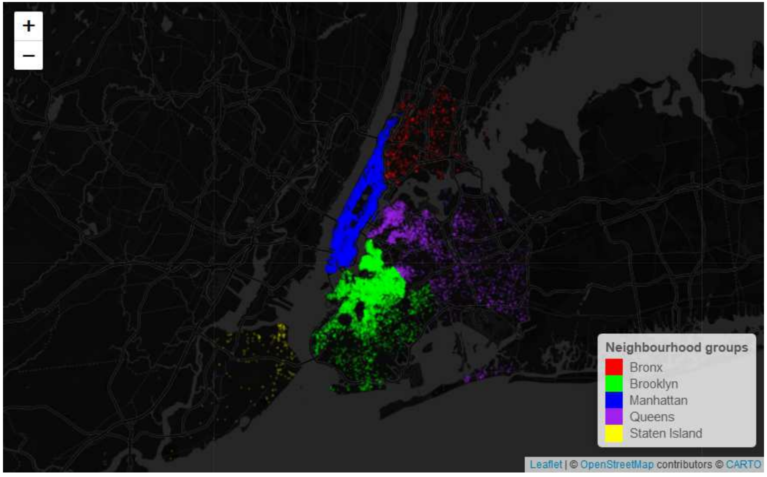
Figure 5 and 6 show that, in terms of sheer number, Manhattan, Brooklyn and Queens were three boroughs where the number of listings increased steeply. However, in terms of percentage increase, Bronx and Staten Island rank high with each of them having recorded 49% and 36% increase from 2018 to 2020 followed by Queens, Brooklyn and Manhattan.

Figure 7: Map Visualization of Listings in 2017 with Colors Corresponding to Different Boroughs



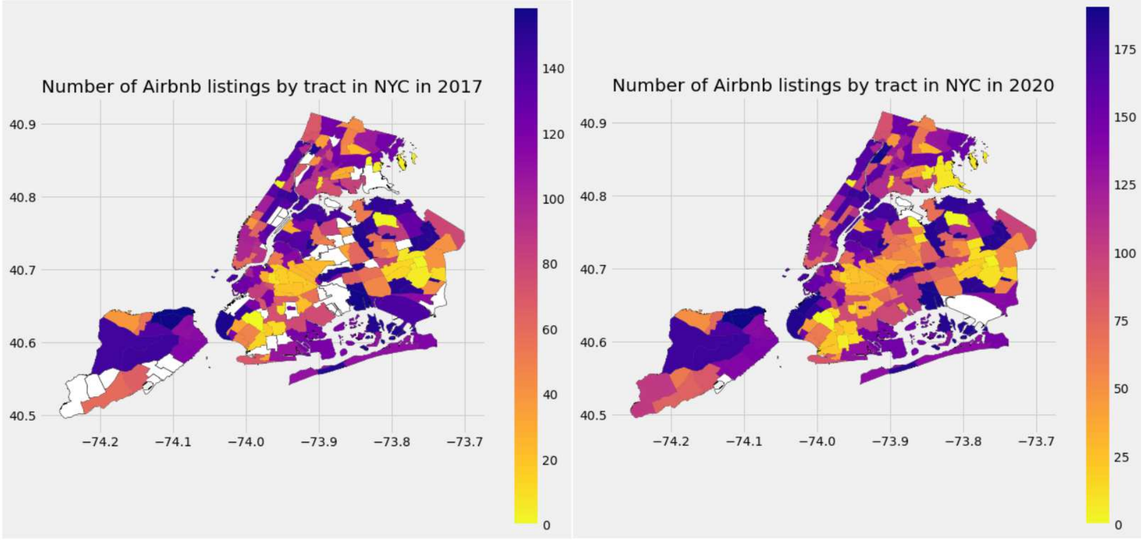
Note: Airbnb Listings visualization in 2017. Different colors represent different boroughs/counties.

Figure 8: Map Visualization of Listings in 2020 with Colors Corresponding to Different Boroughs



Note: Airbnb Listings visualization in 2020. Different colors represent different boroughs/counties. It can be seen that new areas, especially in Manhattan, Brooklyn and Queens, are populated with dots which represent new listings that were added over the past 3 years since 2017.

Figure 9/10: Heat Map for Number of Listings by Neighborhoods/Tracts in 2017 and 2020

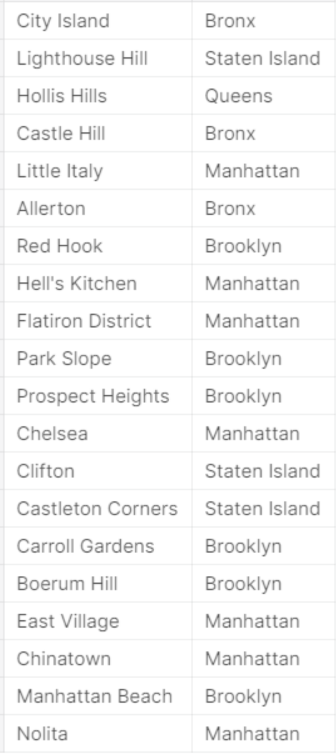


From [Figures 8-10], I discovered that the number of Airbnb listings has been increasing the most in areas of Queens, Bronx and Staten Islands that are closest to Manhattan and Brooklyn (e.g. East and South of Williamsburg, Northeast of Staten Island, Harlem etc.). This matches with my previous point that "gentrification tends to spill over to the adjacent region" in section 5.3 when I explained about luxurious listings.

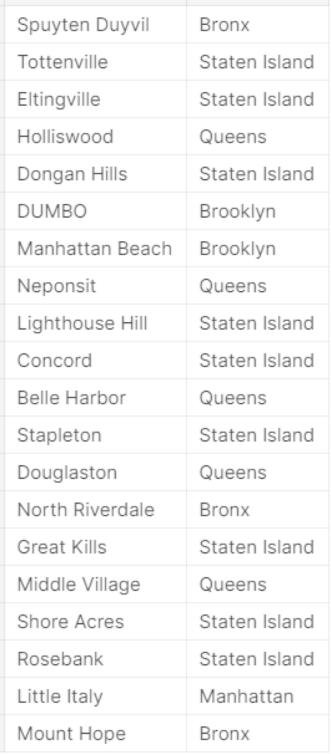
5.5 Minimum Revenue Analysis

I discovered that most of the top 20 neighborhoods with highest minimum revenue in 2017 from [Table 8] correspond to places where gentrification was ongoing or was complete including Chelsea, Hell's Kitchen, Flatiron District, and East Village. This overall supports the rent gap model of Neil Smith (1979) which explains how Airbnb listings can lead to gentrification. From this, I would argue that high minimum revenue can be a warning to ongoing or near future gentrification. In [Table 9], most of the neighborhoods with high minimum revenue are in areas of Queens, Bronx and State Island that are close to Manhattan and Brooklyn. I would argue that these areas have a high probability of gentrification in the near future.

Table 8/9: Top 20 Neighborhoods with highest minimum revenue in 2017 (left) and 2020 (right)



City Island	Bronx
Lighthouse Hill	Staten Island
Hollis Hills	Queens
Castle Hill	Bronx
Little Italy	Manhattan
Allerton	Bronx
Red Hook	Brooklyn
Hell's Kitchen	Manhattan
Flatiron District	Manhattan
Park Slope	Brooklyn
Prospect Heights	Brooklyn
Chelsea	Manhattan
Clifton	Staten Island
Castleton Corners	Staten Island
Carroll Gardens	Brooklyn
Boerum Hill	Brooklyn
East Village	Manhattan
Chinatown	Manhattan
Manhattan Beach	Brooklyn
Nolita	Manhattan



Spuyten Duyvil	Bronx
Tottenville	Staten Island
Eltingville	Staten Island
Holliswood	Queens
Dongan Hills	Staten Island
DUMBO	Brooklyn
Manhattan Beach	Brooklyn
Neponsit	Queens
Lighthouse Hill	Staten Island
Concord	Staten Island
Belle Harbor	Queens
Stapleton	Staten Island
Douglaston	Queens
North Riverdale	Bronx
Great Kills	Staten Island
Middle Village	Queens
Shore Acres	Staten Island
Rosebank	Staten Island
Little Italy	Manhattan
Mount Hope	Bronx

6 Results and Discussion of Predictive Modelling

Table 10: Baseline Performances on Testing Data

	MSE	Median AE	Mean AE	R Squared
Linear Regression	22959.27	30.71	49.62	0.15
Ridge	22959.25	30.70	49.62	0.15
Lasso	22983.24	30.59	49.49	0.15
Elastic Net	24724.06	42.34	58.98	0.08

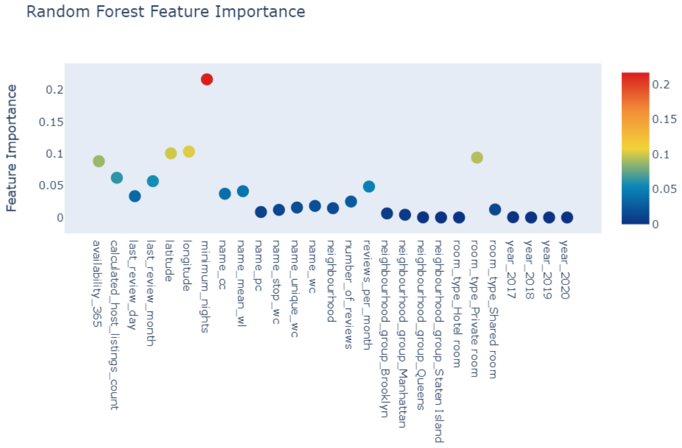
Table 11: Performances on Testing Data of Tree / Forest Derived Algorithms

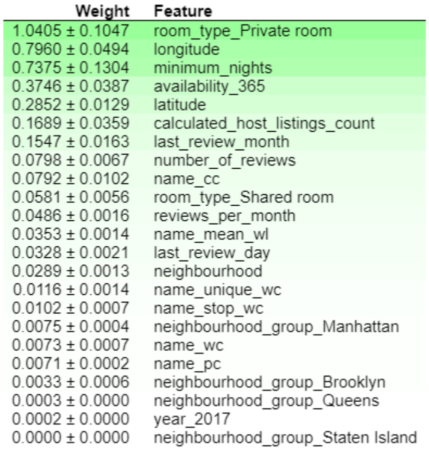
	MSE	Median AE	Mean AE	R Squared
RF (n_estimators = 50)	32.51	0.30	1.37	0.99
GBR (n_estimators = 50)	14002.13	24.75	44.08	0.48
XGBoost	5359.64	22.49	38.30	0.80
LGBM	798.12	11.79	11.76	0.97
RF on features with non-zero RF feature importance (n_estimators = 50)	22.82	0.30	1.34	0.99
RF on features selected by Lasso (n_estimators = 50)	17.96	0.30	1.33	0.99
RF on features with non-zero RF feature importance (n_estimators = 300)	11.34	1.18	0.47	0.99
Bagging + RF on features with non-zero RF feature importance (n_estimators = 300)	352.01	2.94	6.36	0.99

Note: Red indicates the best performing model.

Baseline linear regularized models scored a Mean Absolute Error range of 30-40 and a Mean Absolute Error range of 22900-25000. These performances improved significantly when tree and forest based algorithms were applied combined with bagging and boosting methods. The best performing model was Random Forest applied on the feature-selected dataset using 300 different bagging trees. This has beaten the performances of the best model from the previous literature aforementioned in section 2 for Mean Absolute Error (1.1584) and R-Squared (0.6901). My best model still performs worse than that from previous literature for Mean Squared Error but I was still able to make a major improvement for the Mean Squared Error of around 22900 to 11 in comparison to the baseline model I first created. Moreover, I have beaten Samuel Lam's personal project performance for Median Absolute Error of around $22. Importantly, I was on-par or better in listing price prediction performance than previous literature using smaller number of features.

Figure 11/12: Random Forest Feature Importance (left) and Permutation Importance (right).





Weight	Feature
1.0405 ± 0.1047	room_type_Private room
0.7960 ± 0.0494	longitude
0.7375 ± 0.1304	minimum_nights
0.3746 ± 0.0387	availability_365
0.2852 ± 0.0129	latitude
0.1689 ± 0.0359	calculated_host_listings_count
0.1547 ± 0.0163	last_review_month
0.0798 ± 0.0067	number_of_reviews
0.0792 ± 0.0102	name_cc
0.0581 ± 0.0056	room_type_Shared room
0.0486 ± 0.0016	reviews_per_month
0.0353 ± 0.0014	name_mean_wl
0.0328 ± 0.0021	last_review_day
0.0289 ± 0.0013	neighbourhood
0.0116 ± 0.0014	name_unique_wc
0.0102 ± 0.0007	name_stop_wc
0.0075 ± 0.0004	neighbourhood_group_Manhattan
0.0073 ± 0.0007	name_wc
0.0071 ± 0.0002	name_pc
0.0033 ± 0.0006	neighbourhood_group_Brooklyn
0.0003 ± 0.0000	neighbourhood_group_Queens
0.0002 ± 0.0000	year_2017
0.0000 ± 0.0000	neighbourhood_group_Staten Island

For feature importance, both feature importance function of Random Forest and permutation important point to similar set of features as the "most important features" in predicting the price of an Airbnb listing in New York City. Those were whether a room is a private room or not, number of available days for the next 365 days since the listing was posted on Airbnb, minimum number of nights the guest has to say in that listing and the location information (longitude, latitude). On the other hand, features that were considered to be unimportant were the dummy variables indicating whether the listing was from 2018-2020. Features year_2016 and year_2017 did not have zero scores for feature importance although they were still trivial in the prediction process.

7 Limitations and Future Work

In order to fully gauge the impact of Airbnbs on gentrification, I will have to collect short term rental supply and pricing data from sources including Zillow. It would also be interesting to collect the American Community Survey Data on a neighborhood level to evaluate the impact of Airnbnbs on New York City's various socioeconomic and demographic indicators. In the future, I hope to utilize other features including the text data from user reviews and create neural networks to improve the predictive power of the my models. For enhanced feature interpretability, I can consider using another feature importance method called "SHAP" to capture the directionality of the impact of respective features in predicting the target variable.

8 References

Neighborhood Tabulation Areas (NTA) data, New York City Open Data
https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq

A Year Later: Airbnb as a Racial Gentrification Tool, Inside Airbnb
http://insideairbnb.com/face-of-airbnb-nyc/a-year-later-airbnb-as-racial-gentrification-tool.html

The High Cost of Short-Term Rentals in New York City by David Wachsmuth, David Chaney, Danielle Kerrigan, Andrea Shillolo, and Robin Basalaev-Binder, Urban Politics and Governance Research Group, McGill University
http://www.sharebetter.org/wp-content/uploads/2018/01/High-Cost-Short-Term-Rentals.pdf

A socioeconomic analysis of Airbnb in New York City by Lajos Boros, Gabor Dudas, Gyorgy Vida, and Tamas Kovalcsik
https://www.researchgate.net/publication/320443736_A_socio-economic_analysis_of_Airbnb_in_New_York_City

Then as Now — New York's Shifting Ethnic Mosaic by Ford Fessenden and Sam Roberts, NYTimes
http://archive.nytimes.com/www.nytimes.com/interactive/2011/01/23/nyregion/20110123-nyc-ethnic-neighborhoods-map.html?_r=0

A story map: AirBnB and Gentrification, ArcGis
https://www.arcgis.com/apps/MapJournal/index.html?appid=57dfff16e5fd4eabb0d187043eaf695a

Research: When Airbnb Listings in a City Increase, So Do Rent Prices by Kyle Barron, Edward Kung and Davide Proserpio, Harvard Business Review, 4/17/2019
https://hbr.org/2019/04/research-when-airbnb-listings-in-a-city-increase-so-do-rent-prices

Airbnb and Rent Gap: Gentrification Through the Sharing Economy by David Wachsmuth and Alexander Weisler
https://journals.sagepub.com/doi/full/10.1177/0308518X18778038

Airbnb Price Prediction Using Machine Learning and Sentiment Analysis by Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei
https://arxiv.org/abs/1907.12665

Airbnb Pricing Predictions, Samuel Lam
https://airbnb-pricing-prediction.herokuapp.com/