APMA E4990: Modeling Social Data

Columbia University, Spring 2017

airbnb Hosts in New York City

Final Project

Due: Friday, May 5

Tejas Dharamsi td2520
Vedant Dharnidharka vd2334
Molly Hanson meh2243

Introduction

Airbnb, an online marketplace that allows hosts to rent out their vacant living space to short-term renters, has been considered as a disruptive technology (Alton, 2016). By Airbnb providing consumers with a service that did not previously exist, there has been a major shift of power within the hotel industry specifically and the company has received significant scrutiny from various local parties. The focus of the following report is to analyze Airbnb hosts in New York City with respect to two lens: how can hosts benefit by renting their space using Airbnb, and what are they at risk of by using this home-sharing platform.

Host Benefits

Alton (2016) highlights a preference among today's consumers to seek out a product or service from a person, rather than from a large corporation. Airbnb does just that by connecting individuals with each other, and their Superhost program incentivizes this further by providing hosts with extra perks such as travel coupons, product exclusivity and priority support. Airbnb (n.d.) categorizes Superhosts as extraordinary hosts who have met the following benchmarks in the following categories: sufficient completed listings, high response rate, majority 5-star ratings, high review rate, and commitment to following through with reservations.

The main goal here is to determine if a model based on these listed Superhost benchmarks can predict a host's Superhost status, or if there are other variables that influence this status. This is an important question, as it will help validate if the program is as transparent as it seems, or if there are other variables (perhaps even subjective ones) that affect one's Superhost status. Moreover, interested hosts can use this model to outline what they need to do to increase the likelihood of becoming a Superhost.

Host Risks

As briefly noted above, alongside disruption comes scrutiny and backlash. In New York City especially, Benner (2016) lists numerous parties in opposition of Airbnb including hotel unions, landlord and tenant groups, affordable housing advocates and politicians from both parties. New York state law prohibits individuals for renting out their apartments for fewer than 30 days, but as of now, it is not illegal to advertise short-term home rentals (Nayak, 2016). Those in opposition argue that platforms like Airbnb have worsened the affordable housing issue in cities like New York: affordable housing advocate Neal Kwarta explained "Airbnb is dominated by commercial operators with multiple listings who are stealing our supply of affordable housing" (Benner, 2016).

The goal of this part of the report is to assess the concentration and characteristics of hosts with multiple listings, as well as zoom in on this from a neighborhood perspective in connection with affordable housing concerns. This is important as Airbnb has a "One Host, One Home" policy, thus their terms prohibit hosts to list multiple apartments. Moreover, as New York officials consider banning advertising of short-term apartment listings and imposing fines those who continue to do so, New York City could set a precedent for other localities, especially those with housing shortages, consequently impacting Airbnb's future global operations.

Part I: Can a model predict Superhost status?

Introduction

Airbnb launched a Superhost program to incentivize hosts to be superb hosts, as renters are less likely to continue using Airbnb if they run into issues with their host. In order to become a Superhost, Airbnb (n.d.) lists basic criteria that must be met, over the course of a year:

- Hosted at least 10 trips
- Maintained a 90% response rate or higher
- Received a 5-star review at least 80% of the time you've been reviewed, as long as at least
- Half of the guests who stayed with you left a review
- Completed each of your confirmed reservations without canceling

The goal our model is to predict whether a host is a Superhost or not. First, this will be done using a variety of features in the dataset. Next, we will run a model only using selected features that reflect the Superhost criteria above. The overall goal of this comparison is to see if the traits Airbnb claim to be all that is required to be a Superhost are sufficient in predicting the status, or if there are other influential features, such as number of bedrooms and availability rate.

Data Source

The data for this report was obtained from insideairbnb.com. *Insideairbnb* sources their data from publically available information on the Airbnb website directly. *Insideairbnb* provides data based on Listings, Calendar and Reviews, for most of the major cities where Airbnb operates. The below analysis utilizes the New York City Listings dataset, which was last scraped on April 2, 2017.

Data Cleaning + Sanity Checks

Prior to delving into the modeling task, it was important to assess the data quality and clean the data appropriately. Initially, the dataset contained over 95 features describing more than 46,000 listing. Large portions of these features were irrelevant to the modeling task outlined above. Therefore, roughly 73 features were removed from the dataset, with the most relevant features remaining.

Newer listings with fewer visits did not have sufficient data among most of the remaining features. Several relevant parameters such as response rate, response time and review_scores could not be assigned values during pre-processing to newer listings with insufficient visits. Due to this, roughly 13,000 data points referencing relatively new listings were removed due to insufficient data. A final pre-processing step included creating one-hot encoding for features with non-numeric values (categorical variables), which could be converted into different classes, such as instant_bookable, response_time and room_type.

It is important to mention that even after the removal of these newer listings, the data was highly imbalanced: out of 22,891 listings, only 3,320 were listings by a Superhost. In the subsequent section outlining the modeling process, there is a detailed discussion of how this imbalance was approached.

Modeling Approach

The modeling task at hand is a classification problem. This is the natural way to frame this task as the goal is to predict the class of a host as a Superhost (1) or not (0). Several classification models were considered throughout the process, with the initial being a probabilistic model: Logistic Regression. Preliminary results using Logistic Regression did not generate strong results; accuracy was just at par with the baseline for predicting 0 (non-Superhost), thus we did not proceed with this model and considered other modeling approaches.

Next, a classification model based on Random Forest Regression was considered. After the implementation of a basic Random Forest Regression with no hyperparameter tuning, the results were superior to those from Logistic Regression, however results were still below the baseline for predicting 0. This was, as briefly introduced earlier, largely related to the imbalance in data. Imbalanced data can be dealt with in multiple ways. The two approaches considered – under-sampling and class weight balance – all gave varying results.

After evaluating our model with both different parameters and different forms of sampling to offset class imbalance effects, a Random Forest model using balanced sampling realized 88% accuracy, which is comfortably above the baseline for predicting 0 for all cases.

Model Performance

Each model below is assessed based on accuracy, precision and recall, and AUC score, and as this is a classification task, a confusion matrix is also included for each model. The following outlines the results for each model, as well as the hyperparameters used. Models 1 and 2 utilize the entire cleaned subset of variables, while Model 3 only uses the variables considered as a proxy to the Superhost traits.

Model 1: Class-Weight Balanced

This model was built using Scikit-Learn's RandomForest Classification Algorithm. Due to the high imbalance of the data points not being a Superhost, the "class_weight=balanced" parameter was utilized to give more weight to the class with fewer data points. The n_estimator (number of trees in the forest) parameter was varied between 50 to 300 and max_depth (maximum depth of the tree) between 10 to 30 to find the optimal parameter value. This is an instance of complexity control, whereby the model was tuned using several values for n_estimator and the best model parameter were selected based on number of misclassified observations (in this case, n_estimator = 200, max_depth =16). Next, the data was split into 75/25 train/test sets, and 5-fold cross validation was performed on the train dataset. Training data realized 87.56% accuracy.

Test results (Model 1): 88.1% Accuracy, 0.887 AUC Score

	PRECISION	**RECALL**	**F1-SCORE**
0	0.94	0.92	0.93
1	0.57	0.65	0.61
AVG	0.89	0.88	0.88

Confusion Matrix:

	0	*1*
0	4,505	399
1	284	535

Model 2: Under-sampling

When there is notable imbalance in the classification task, under-sampling is a common approach used to achieve a balanced dataset. This model again used scikit-learn's Random Forest

Classifier Algorithm, however random under-sampling was performed. In this scenario, 3,000 data points were randomly picked from each class: Superhost (1) and non-Superhost (0). Next, this dataset containing 6,000 observations was divided into 75/25 train/test sets. Once again, using 5-fold cross validation on the training set, the model obtained 79.93% accuracy.

Test results (Model 2): 79.4% Accuracy, 0.883 AUC Score

	PRECISION	**RECALL**	**F1-SCORE**
0	0.79	0.80	0.80
1	0.80	0.79	0.79
AVG	0.79	0.79	0.79

Confusion Matrix:

	0	*1*
0	600	150
1	159	591

Model 3: Under-sampling on 4 “Superhost” traits

In this version of the model we repeat the previous model 2 with four closest features that resemble the criterions stated by Airbnb for Superhost:

- host_response_rate
- host_response_time
- review_score_rating
- number_of_reviews

This model gives nearly similar results as compared to Model 2, but with only 4 features. This suggests these features are the most important features in classifying hosts as super hosts, which is aligned with the Superhost program traits indicated on the company’s website. Once again, using 5-fold cross validation on the training set, the model obtained 80.13% accuracy.

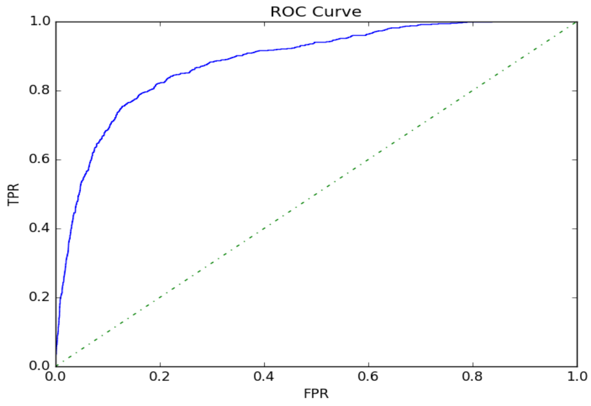
Test results (Model 3): 78.86% Accuracy, 0.8665 AUC Score

	PRECISION	**RECALL**	**F1-SCORE**
0	0.77	0.83	0.80
1	0.81	0.75	0.78
AVG	0.79	0.79	0.79

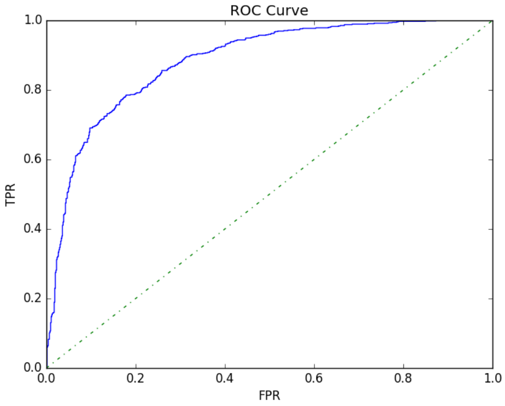
Confusion Matrix:

	0	*1*
0	622	128
1	189	561

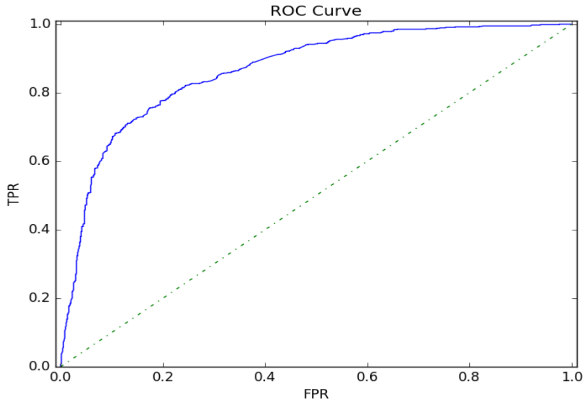
ROC Curves



Model 1: 0.8844 AUC Score



Model 2: 0.883 AUC Score



Model 3: 0.844 AUC Score

Runtime Complexity and Scalability

Random Forests are an ensemble model of decision trees. Time complexity for building a complete, unpruned decision tree is $O(v \times n \log(n))$, where n is the number of records and v is the number of variables/attributes.

The two main parameters one must define when building Random Forests are: the number of trees to be built (assumed to be, $ntree$) and how many variables should be at sampled at each node (assumed to be, m). Since we would only use m variables at each node, the complexity to build one tree is $O(m \times n \log(n))$. Expanding this to build a random forest of $ntree$ trees, the complexity becomes $O(\, ntree \times m \times n \log(n))$. This is the worst case scenario complexity, which assumes the depth of the tree to be $O(\log{(n)})$. However, in most cases, the building of a tree stops much before this depth, although this is hard to estimate.

Related to scalability, the framework of this model can be expanded to evaluate other cities, and over longer time frames. Nonetheless, after expanding the dataset in a vast manner, the model may have to run parallel on multiple machines to help with efficiency. Adjustments to the model may be necessary if expanding the computation to multiple computers.

Summary

As mentioned above, the goal of this modeling task was to verify the features which are taken into consideration while determining the Superhost status of the host, as mentioned on the Airbnb website. The above RandomClassifier Model with under-sampling (Model 2) overcame the imbalance on the dataset and achieved results well beyond the 50% baseline. However, similar results from Model 3 indicate that the Superhost traits are sufficient in predicting the Superhost status, and including other variables does not improve predictions by a significant margin. Thus, in the interest of model complexity, it is recommended to use Model 3. In all, this confirms the transparency of the Superhost program, and those looking to become Superhosts should examine Model 3 to increase the likelihood of gaining Superhost status.

With that being said, there is room for model improvement. Notably, not all of the traits were represented in the dataset, so including such information could further the success of this model.

It is important to note that one of our initial ideas was to incorporate the Reviews data as a proxy for the 80% positive feedback trait. However, we came across two issues with the review structure and available data set. Firstly, renters are not obliged to write a review, so biases in the data set are inevitable. Moreover, the data set does not have a rating score associated with each review, so there is no "ground truth" to use in training a model. Although there is no way around the first issue, we experimented with a prepackaged Python library to extract sentiment. However, we were not confident in the results derived from this library thus it was omitted from the analysis and modeling task.

Part II: Are multi-listers a concern in New York City?

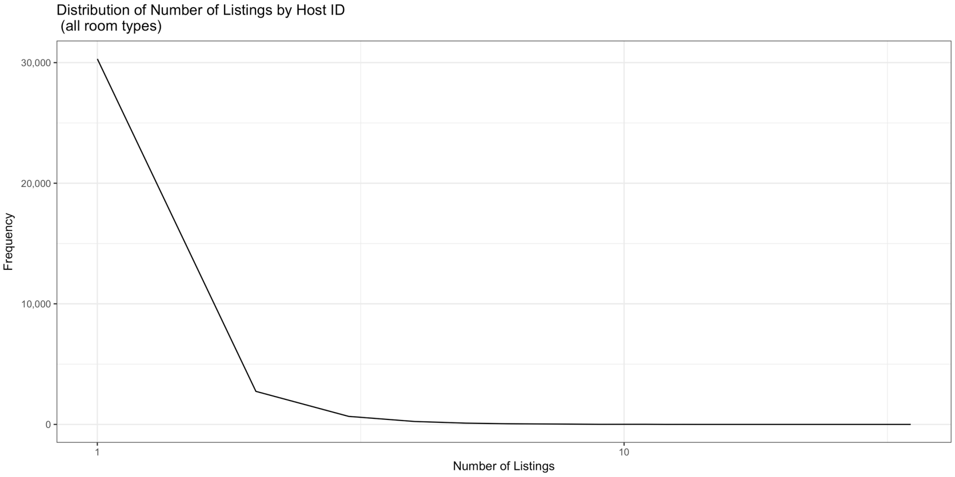
Introduction

Airbnb notably disrupted the tourism and hospitality industry by providing tourists with lower cost alternatives to hotels. Additionally, renters reap the benefits of additional income as Airbnb "helps residents defray high rents" (Nayak, 2016). While controversy persists regarding the forgone hotel taxes by tourists residing in Airbnb's rather than hotels, the hotel industry and regulators also argue that Airbnb is "leading to a proliferation of homes that essentially function as illegal hotels in violation of zoning laws, safety codes and other requirements" (Nayak, 2016).

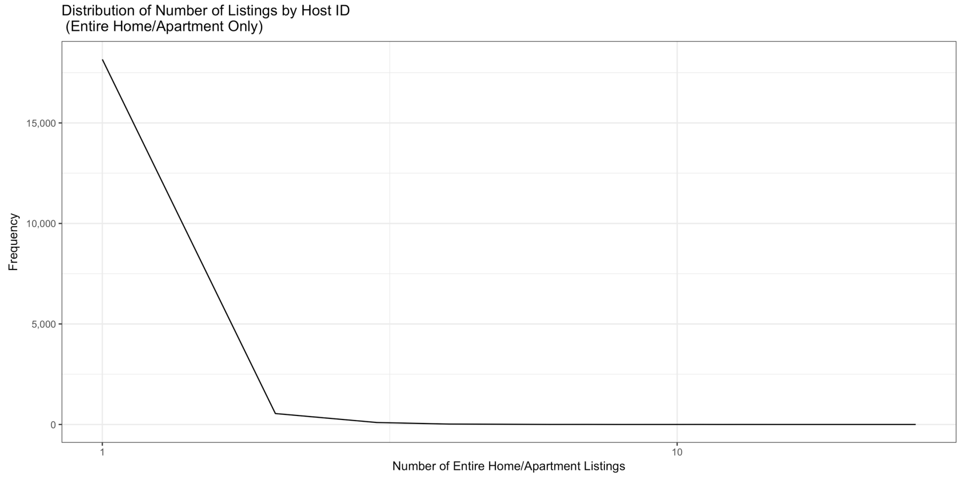
Despite the importance of these hospitality and tourism concerns, the following analysis focuses on a greater social issue: Airbnb's disruption of the housing market. Those in opposition claim Airbnb diverts units out of the housing market, however it is hard to argue that under Airbnb's "One Host, One Home" policy, the housing market is greatly at risk. That being said, commercial operators with multiple listings do pose a threat to the housing market, and Airbnb supports the "crack down on bad actors" (Walters, 2017).

Concentration of Multi-Listers

Airbnb consistently publicizes their continuous efforts to remove multi-listing hosts from their platform. While the majority of hosts only have one listing (under a given host ID), of the 34,223 unique host IDs in this dataset, 11.4% have multiple listing. It is important to note this distribution includes all room types: private room, entire home or apartment, and shared room.



Affordable housing advocates are primarily concerned with renting entire apartments. Thus, by filtering the dataset to only include entire apartments and subsequently assessing the distribution of number of listings, only 3.6% of hosts who list an entire home or apartment have multiple entire home listings. This observation is in line with the company's public statements which estimate "96% of New Yorkers renting out an entire unit were offering their own home" (Walters, 2017).

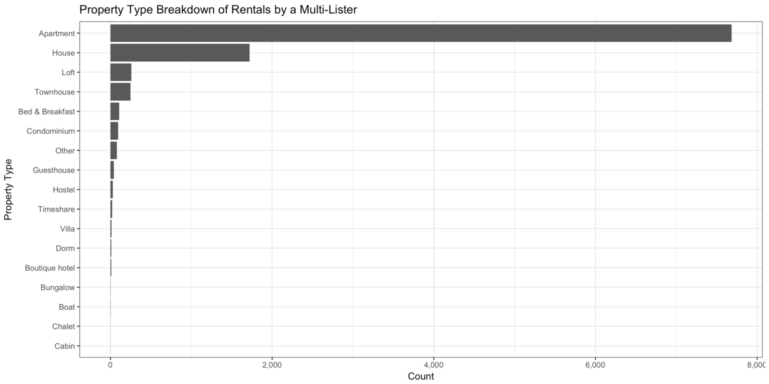


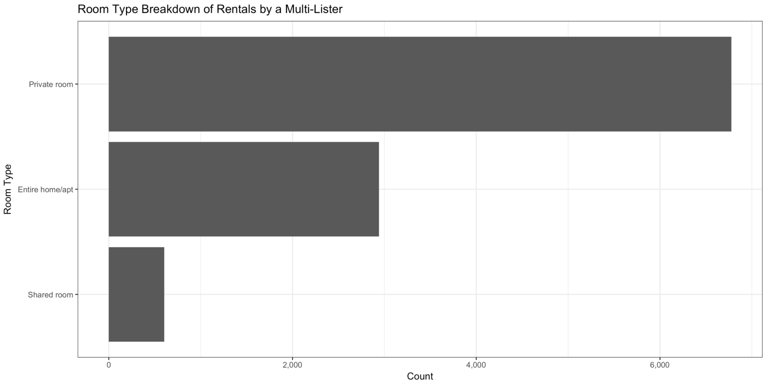
Also, the listings dataset contains the variable "host listing count". When comparing this variable to the count of grouping by Host IDs, 6.7% of the host IDs had inconsistent counts. More notably, the ranges of these variables were substantially different: the range of number of listings when using the group by function is 1 to 35, whereas the provided variable ranges from 0 to 856 listings, which is clearly invalid. It is important to note that all analyses were completed using the calculated number of listings field, rather than the provided "host listing count" variable.

Who are Multi-Listers?

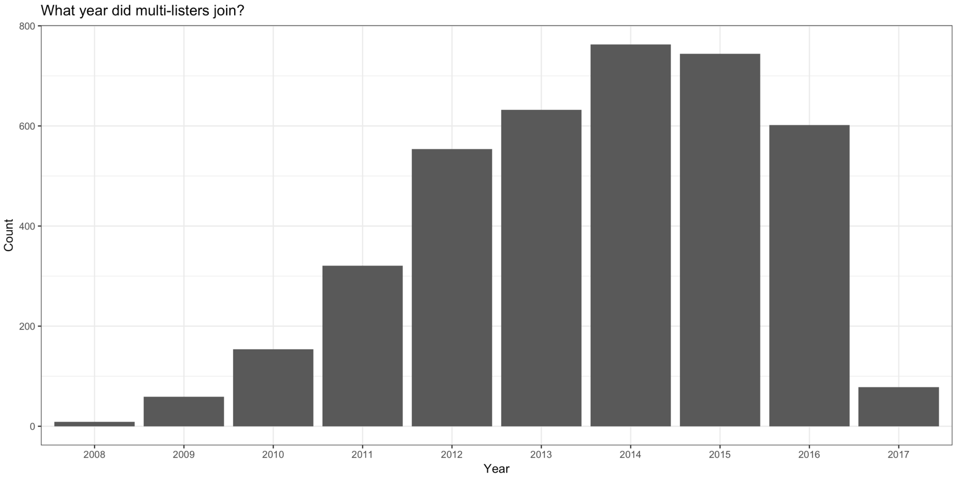
While it is harder to directly link renting a short-term private or shared room with the affordable housing market, and such multi-listers may not be forthright violating the "One Host, One Home" policy, it would not be valid to say there is no relationship. The following analysis which profiles who multi-listers are, considers those who list any room type.

Looking only at New York City listings posted by a multi-lister, unsurprisingly the majority of rentals are apartments and houses. To this end, 66% of rentals listed by a multi-host are a private room, and 28% are entire homes or apartments.





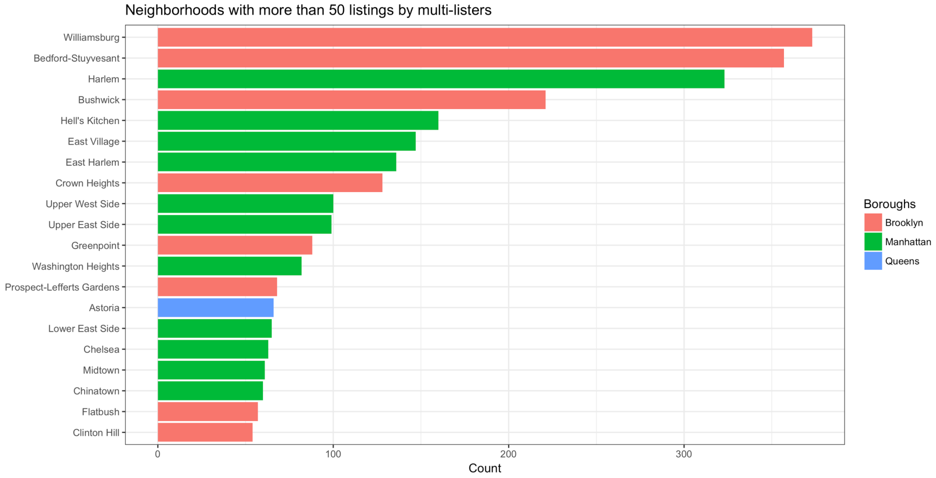
The join year of hosts who are currently multi-listers has not been consistent over time. In fact, fewer current multi-listers in New York City joined 2015 or 2016, compared to 2014. This may suggest that newer hosts are more respectful of Airbnb policy, or newer hosts that list multiple homes have already been removed from the platform in various purges.



Lastly, as stressed in the above section, Superhosts are rewarded hosts based on various criteria. As multi-listers, particularly those listing multiple entire homes, are violating Airbnb terms, investigation into the breakdown of multi-listing Superhosts seemed worthwhile. Analysis showed that 13% of multi-listers are Superhosts, which seems high as hosts are intended to rent only their primary residence at a time.

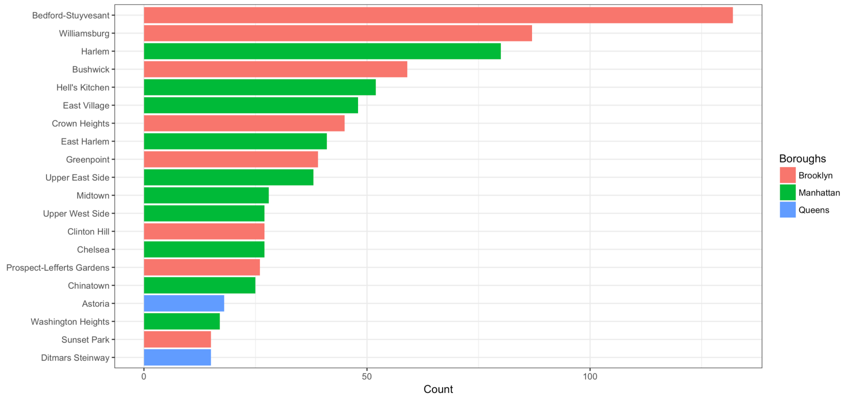
Location of Multi-Listers' Listings

As many have expressed concern over Airbnb taking affordable housing away from New Yorkers, the following assesses listings by multi-listers, by neighborhood. Accounting for all room types, 19 of the 20 neighborhoods with the most listings by multi-listers are in Brooklyn or Manhattan.



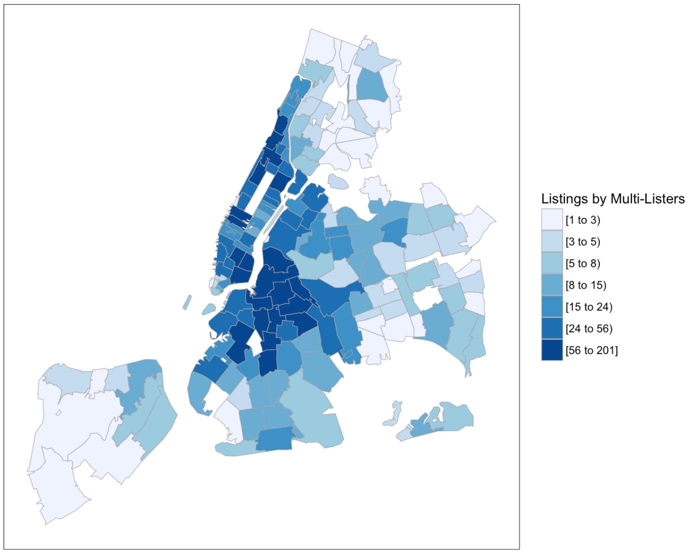
However, as affordable housing advocates are largely concerned with the short-term rental of entire apartments, the same analysis was completed with only this room type in mind. The results are somewhat similar, but to a much lower scale as entire apartments represent less than 30% of all listings by multi-listers. Bedford-Stuyvesant is clearly an area of high concern related to housing affordability; 132 full apartment or homes in this neighborhood are currently listed on Airbnb by multi-listers. This may not be a large number on its own, but this adds up over time and over neighborhoods, and extending this view globally, over other cities around the world.

Neighborhoods with more than 15 entire home/apartment listings
by multi-listers

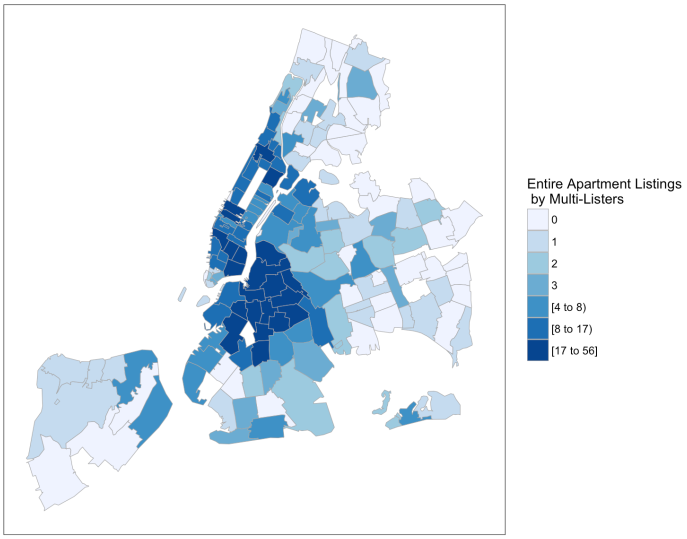


Looking at these distributions on a map, it is clear that the Brooklyn neighborhoods nearest to Manhattan, and those in Manhattan, have the highest concentration of listings by multi-listers, regardless of room type.

New York City Airbnb Listings by Multi-Listers



New York City Airbnb Entire Home/Apartment Listings by Multi-Listers



As noted above, one of the main pain points related to Airbnb is the perception that commercial operators with multiple listings are stealing the supply of affordable housing in the city. To investigate the impact of this, the following analysis seeks to determine if the full apartments that a multi-lister hosts are in the same building. The idea here is that perhaps a commercial operator has several units in one building, and rather than renting these apartments out to individuals who want to live in New York, the hosts default to Airbnb, in turn worsening the city's housing shortage and affordability concerns.

To calculate closeness of entire apartments listed by the same host, first the dataset was grouped by the Host ID, and then for each grouped Host ID, we calculated the mean latitude and longitude of their listings. Next, the listing's actual latitude and longitude was subtracted from the respective mean. A simplifying assumption was made: if both the absolute difference of the latitude and longitude from the mean were less than 0.001, it was concluded these listings were in the same building. While this generalization is not perfect, it is possible there is calculation error in the provided latitude and longitude data and the provided values are not perfectly accurate. Nonetheless, this can provide an overall idea for the volume of multi-listers who list multiple full homes in a close vicinity.

In all, of the entire homes or apartments listed by a multi-lister, based on the above assumptions of closeness, 34% of such listings are in the same building. Moreover, of all the entire homes and apartments listed on Airbnb in New York City at the time when this data was scrapped, 5% of these are by a multi-lister and in the same building. When focusing on the number of "bad actors", there are 463 hosts who are should be removed promptly from Airbnb as these individuals have multiple listings of entire homes and apartments that we consider to be in the same building.

Summary

The above analysis suggests there are a handful of hosts in violation of the "One Host, One Home" policy: 3,917 hosts have multi-listings, and 694 hosts list multiple entire apartments. While these numbers represent a relatively small percentage of total hosts, parties in opposition still see this as an opportunity to target Airbnb. As mentioned, New York prohibits rentals for less than 30 days when the resident is not present, but as of now, advertising such rentals is not explicitly legal; however newly proposed legislation would make this a finable act (Nayak, 2016).

Walters (2017) confirms that regulators are not concerned with "the little guy", and with limited resources the focus is on "serial violators who rent out one or two or more units beyond their own personal unit". However, this proposed law does not protect individuals who use Airbnb for its purpose and some extra income. By passing such legislation, New York City would be the first to ban ads for short-term rentals on home-sharing sites (Nayak, 2016), which may have greater global effects.

Looking Ahead

Through our model, we were able to assess the applicability of the metrics and the benchmarks set by Airbnb with respect to the Superhost tag. When considering areas of future exploration, it would be interesting to consider the reviews dataset, and see how well hosts take into consideration reviews – both positive and negative. Based on these reviews, an interesting analysis would be to assess whether after negative comments there is any improvement in next guest's experience at the listing. With this in mind, we can categorize hosts as those who take reviews seriously, and thus refine the feedback model.

In all, the previous analysis of multi-listers confirm Airbnb's publicized statements regarding hosts listing multiple homes. However, with much pressure from politicians and affordable housing advocates on the company's New York operations, Airbnb will have to take quick measures to eliminate multi-listers all together. If this is not completed to the standards of those in opposition, it is possible regulations banning advertising on home-sharing sites will become a reality, which may have rippling effects to other cities. It is somewhat puzzling that a technology company has not hard-coded measures to block hosts from creating multiple listings. Said (2016) reports that Airbnb is updating its technology in such a way to bar San Francisco hosts who manage multiple listings. However, it would seem this automation would be a top priority for the company, especially in New York City where prohibitive laws against the company's services are in the process of being passed.

Sources

Airbnb. (n.d.). Superhost. Retrieved from https://www.airbnb.ca/superhost

Alton, L. (2016, April 11). How Purple, Uber and Airbnb are disrupting and redefining old industries. Entrepreneur. Retrieved from https://www.entrepreneur.com/article/273650

Benner, K. (2016, Oct 19). Airbnb proposes cracking down on New York City hosts. The New York Times. Retrieved from https://www.nytimes.com/2016/10/20/technology/airbnb-proposes-cracking-down-on-new-york-city-hosts.html?_r=0

Inside Airbnb. (2017, April 2). Get the Data, New York City, New York, United States. Retrieved from http://insideairbnb.com

Nayak, M. (2016, June 20). New York bill would ban Airbnb listings for some short-term rentals. Reuters. Retrieved from http://www.reuters.com/article/us-new-york-airbnb-idUSKCN0Z62M2

Said, C. (2016, Oct 18). Airbnb bans hosts with multiple listing in SF. SF Gate. Retrieved from http://www.sfgate.com/business/article/Airbnb-bans-hosts-with-multiple-listings-in-SF-9982303.php

Walters, J. (2017, Feb 12). Something in the Airbnb: hosts anxious as New York begins crackdown. The Guardian. Retrieved from https://www.theguardian.com/technology/2017/feb/12/airbnb-hosts-new-york-fines-government-illegal

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.