A Quantitative Analysis on New York City Airbnb Open Data

Yunxin Wang

Introduction

Airbnb, founded in 2008, has enabled guests to book private rooms, homes and apartments of hosts for short-term rental stays. As a new form of rental service, Airbnb has expanded to many local markets around the world while raising questions from residents and policymakers. Among all the controversies, the greatest concerns about Airbnb are its potential negative impacts on local housing availability and prices, local government tax returns, and the quality of life in residential neighborhoods.[1] This report will conduct a data-driven research on these three questions and provide recommendations based on the analysis and results.

The rationales behind these three concerns are as follows. First, Airbnb is believed to reduce the availability of local housings, thus increasing the long-term housing price, because it encourages high- and middle- income groups to buy residences to rent them out. This increases demand for housing, causes prices to climb, profits people with more than one property, while prioritizing travelers over locals.[2] Second, Airbnb reduces reliable lodging taxes for governments, because Airbnb not only contributes little in tax revenue but also cuts taxes from hotels by taking over their customers.[3] Finally, Airbnb brings up concerns from local residents because short-term renters are more likely to create negative externalities, such as noises and trashes.

This report aims to make recommendations on these three concerns, including the impact of Airbnb on local housing availability and prices, tax revenue, and the geographical distribution of Airbnb among local neighborhoods. The primary audiences that the report hopes to communicate with are policymakers in the New York City government. They could use the information from this project because it provides data-based evidences to support three policy decisions.

First, should the government regulate the number of listings that each host should own? This decision should be made based on the impact of Airbnb on local housing availability. If the data shows that 1) a large proportion of Airbnb hosts have more than one listing, 2) most listings are entire homes, and 3) the listings are available for most times of the year, then homeowners are believed to purchase housings purely for renting revenue. This will reduce the local housing availability and drive up housing price, which requires regulations from the local government. Second, if the government would impose taxes on Airbnb, how much tax revenue will generate from Airbnb based on different tax rates and boroughs? Third, this project will recommend the city government to release the geographical distribution map of Airbnb listings to the public.

---

[1] Bivens, J. (2019). The economic costs and benefits of Airbnb. *Economic Policy Institute.* Retrieved from https://www.epi.org/publication/the-economic-costs-and-benefits-of-airbnb-no-reason-for-local-policymakers-to-let-airbnb-bypass-tax-or-regulatory-obligations/

[2] Carey, I. (2019). Airbnb isn't going anywhere. So why aren't cities regulating it more? *Skift*. Retrieved from https://skift.com/2019/02/12/airbnb-isnt-going-anywhere-so-why-arent-cities-regulating-it-more/

[3] Bivens, J. (2019). The economic costs and benefits of Airbnb. *Economic Policy Institute.* Retrieved from https://www.epi.org/publication/the-economic-costs-and-benefits-of-airbnb-no-reason-for-local-policymakers-to-let-airbnb-bypass-tax-or-regulatory-obligations/

The structure of this report is as follows. I will first provide the motivation and background of the problem areas. Then I will introduce the dataset used for this report. After that, I will break down the three problem areas into a number of sub-questions and answer them in turn. For each sub-question, I will describe its rationale, present the result, and provide corresponding analysis. For each problem area, I will synthesize the results of the sub-questions, discuss the final result, and analyze its implications. The sub-questions of each problem area are as follows.

1. About impact on local housing availability and prices:
    1) What percentage of listings are private rooms, entire homes and shared rooms?
    2) What is the number of listings per host on average and its distribution?
       What is the number and percentage of hosts who have more than one listing?
    3) What is the number of days available for booking on average and its distribution?
       What percentage of listings are available for more than 180 days (1/2 of the year)?
    4) What is the average number of listings per host in each borough?
       What is the average number of days available per listing in each borough?

2. About tax revenue:
    1) How much tax revenue in total would a rental tax on New York City Airbnb generate annually based on different tax rates, for example, such as 5%, 7%, and 10%?
    2) What is the average price and total number of reviews for each borough? How would the tax revenue vary with borough respectively with a 5%, 7%, and 10% tax rate?

3. About geographical distribution:
    1) What is the visualization of the geographical distribution of Airbnb listings on the New York City map, based on the provided longitude and latitude? What is the visualization of number of reviews and availability for each listing on the map?

Motivation

These questions are important because the answers to these questions lead to recommendations that serve the public interests. The analysis on local housing availability and prices will protect low-income groups from rising long-term housing prices while preventing the wealthy from exacerbating the unequal distribution of wealth in the society. The inquiry into lodging tax revenues will help government collect tax payments from Airbnb and use the taxes for other public purposes. Finally, the examination of the geographical distribution of Airbnb provides local residents with more information on their neighborhoods because they have the right to know who their neighbors are and what kind of community they are living with.

It is reasonable to answer these questions with quantitative analysis because data presents facts instead of theories and opinions. Furthermore, as explained above, these three questions could be answered with a number of sub-questions that calculates the data based on solid rationale. Therefore, it is both reliable and feasible to use data to answer these questions. Other quantitative research on Airbnb includes “The economic costs and benefits of Airbnb” from Economic Policy Institute, “Regulating Airbnb: how cities deal with perceived negative externalities of short-term rentals” by Nieuwland and Melik, as well as several articles from Towards Data Science.

Dataset

This project will use the New York City Airbnb Open Data on Kaggle. The dataset is originally from Inside Airbnb, an independent, non-commercial open source data tool for Airbnb. The data describes the listing activities and metrics on 48895 Airbnb homes in the New York City. The definitions of the 13 variables included in the dataset are provided below.

id	Listing ID
name	Name of the listing
host_id	Host ID
host_name	Name of the host
neighbourhood_group	Boroughs
neighbourhood	Areas within boroughs
latitude	Latitude coordinates
longtitude	Longitude coordinates
room_type	Listing space type
price	Price in dollars
minimum_nights	Minimum number of nights to stay
number_of_reviews	Number of reviews
last_review	Latest review
reviews_per_month	Number of reviews per month
calculated_host_listings_count	Number of listings per host
availability_365	Number of days when listing is available for booking

To prepare the data for analysis, data cleaning will be performed to deal with the missing data. Unimportant columns for the project, including host_name and last_review, will be dropped. The variable host_name is dropped not only because it is insignificant but also for privacy concerns. The real names of individuals should not be studied or publicized without their consent. Other privacy concerns include exposing the residential location of hosts on the mapping of Airbnb homes. Thus, the mapping will only provide the approximate locations rather than the exact locations of each listing. The copyright of the dataset belongs to Inside Airbnb.[4]

[4] Inside Airbnb. Get the Data. Retrieved from http://insideairbnb.com/get-the-data.html

The Impact of Airbnb on Local Housing

The distribution of room types of Airbnb listings helps examine whether Airbnb has impact on local housing availability and prices. If many listings are entire rooms or apartments, then it is possible that some homeowners purchase rooms and apartments purely for renting. Table 1 displays the number of each room type of Airbnb listings and their percentage. The result demonstrates that 52% of the listings are entire home or apartments, which is relatively high compared with other two types of rooms and increases the likelihood of the proposition above.

TABLE 1. ROOM TYPES OF AIRBNB LISTINGS

room_type	number	percentage
Entire home/apt	25409	51.97%
Private room	22326	45.66%
Shared room	1160	2.37%

However, the distribution of room types is not enough to prove the impact of Airbnb on local housing. It is also necessary to examine the number of listings that each host owns. If many hosts own more than one listing, then it is possible that some homeowners purchase extra apartments purely for renting. Table 2 shows the number of listings per host and their percentage. It was found that the majority of hosts (86.24%) owns only one listing, which indicates that approximately 15% percent of hosts owns more than one listing.

After calculation, it was found that the number of listings per host on average is 1.31. 5154 hosts own more than one listing. Among these hosts, several own abnormally large number of listings. One host owns 327 listings, one owns 232 listing, one owns 121 listings, and another one owns 103 listings. Two hosts own 96 listings. This result shows that some wealthy hosts apparently purchase extra apartments purely for collecting renting revenue.

TABLE 2. NUMBER OF AIRBNB LISTINGS PER HOST

Calculated_host_listings_count	number	percentage
1	32303	86.24%
2	3329	8.89%
3	951	2.54%
…	…	…
96	2	0.0053%
103	1	0.0027%
121	1	0.0027%
232	1	0.0027%
327	1	0.0027%

To prove that some apartments are purchased by hosts entirely for renting, the number of days available for renting of the Airbnb listings also needs to be examined. If the listings are available for booking for more than half the year, then it is believed that these apartments are not lived by the hosts and just for renting. Table 3 shows the top 5 number of days available for booking of the listings. The top 1, surprisingly, is 365 days, with 1295 listings. For the top 2 and 3, 491 listings are available for 364 days and 408 listings are available for one day.

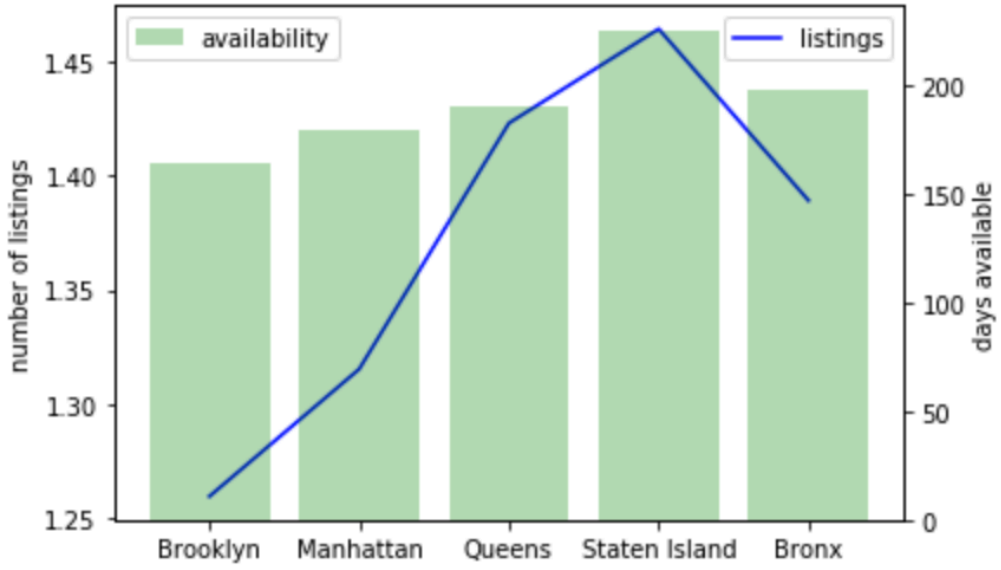
After further calculation, it was found that listings on average are available for 175.83 days per year. 45.8% of listings are available for more than half of the year. The result shows that many listings serve the main purpose of renting instead of also living by the hosts.

TABLE 3. NUMBER OF DAYS AVAILABLE OF AIRBNB LISTINGS

availability_365	number	percentage
365	1295	2.65%
364	491	1.00%
1	408	0.83%
89	361	0.74%
5	340	0.70%

To further understand the impact of Airbnb among different boroughs, the average number of listings per host as well as the average number of days available per listing are plotted in table 4. The result shows that Staten Island has the highest average number of listings per host and the average number of days available per listing, followed by Bronx and Queens.

TABLE 4. AVERAGE HOST LISTINGS COUNT AND AVAILABILITY PER BOROUGH



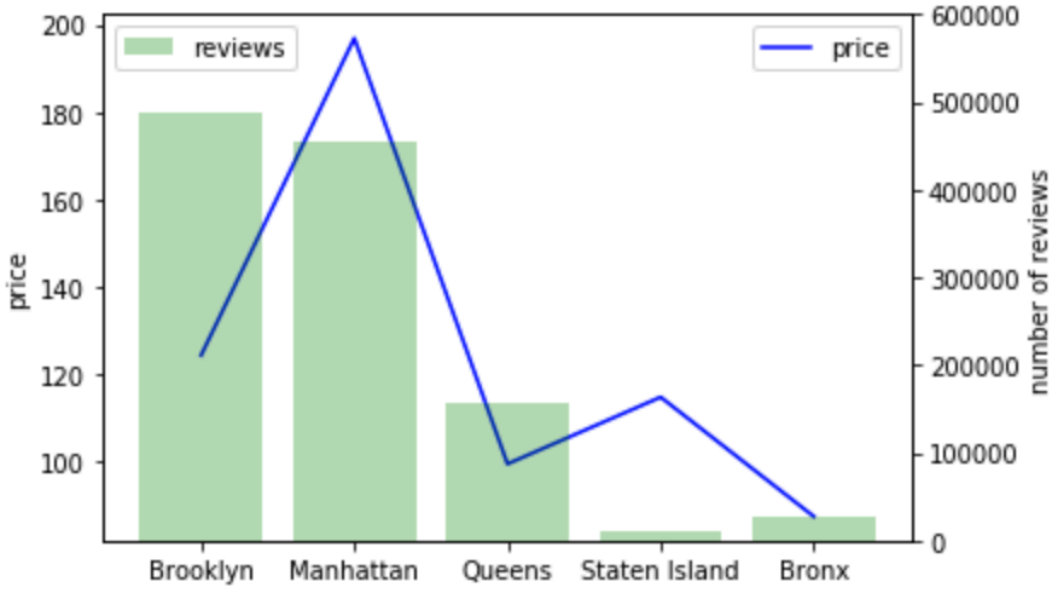
To conclude, the results above regarding room types, number of listings per host, and availability prove that many hosts purchase housings purely for renting. Not only 5154 hosts own more than one listing, some of the hosts even own as many as 327 listings. Furthermore, almost half of the listings are available for more than half of the year. This is problematic because as the middle- and high-income groups buy housings to rent them out, the availability of local housings will decrease and the demand for housings will increase, which drives up the long-term housing price. As a result, the low-income group will bear more pressure in terms of renting and purchasing housings. The government is recommended to pass regulations on the number of listings that each host should own and the number of days available for booking for each listing.

Airbnb and Tax Revenue

The yearly tax revenue from Airbnb is calculated using the function (price * total number of stayed nights * tax rate). However, since the dataset does not have information on total number of stayed nights in Airbnb homes, we need to calculate it using (total number of times of stay * number of nights per stay). Since the dataset also does not have these two variables, we need to estimate the number of stays with total number of reviews and estimate the number of nights per stay with minimum number of nights for staying. Therefore, the final function for calculating tax revenue on each listing is (price * number of reviews * minimum number of nights to stay * tax rate). It is important to acknowledge that this function underestimates the amount of tax revenues on Airbnb, because some customers do not write reviews and some customers stay more than minimum number of nights. However, this function is the best we can do from this dataset.

The annual tax revenue from Airbnb is calculated using this function based on a 5%, 7%, and 10% tax rate. After calculation, the result shows that the annual tax revenue from Airbnb would be $32,505,460 if the tax rate is 5%, $45,507,644 if the tax rate is 7%, and $65,010,920 if the tax rate is 10%. The government is advised to choose the appropriate tax rate based on the result.

TABLE 5. AVERAGE PRICE AND TOTAL NUMBER OF REVIEWS PER BOROUGH



Since tax revenue varies with price and number of reviews, it would be reasonable to analyze how they differ in each borough and as a result, how would the tax revenue vary with borough. Table 5 shows the average price per listing and the total number of reviews in each borough. In terms of price, Manhattan is much more expensive than other four boroughs, with an average price of $197 per listing per night. Brooklyn, Queens, and Staten Island have similar prices around $100. Bronx is the cheapest borough, with an average price of $87 per listing per night. Regarding the total number of reviews, Brooklyn has the most reviews around 486574, followed by Manhattan with 454569. For the rest three boroughs, Queens has 156950 reviews and Bronx has 28371 reviews. Staten Island, with 11541, has the least number of reviews.

Table 6 displays the tax revenue on Airbnb from each borough based on a 5%, 7%, and 10% tax rate. All in all, Manhattan and Brooklyn return the most tax revenue, followed by queens. Bronx and Staten Island return much less tax revenue. The government is recommended to take these factors into account when deciding on the tax rate of Airbnb.

TABLE 6. TAX REVENUE FROM AIRBNB PER BOROUGH

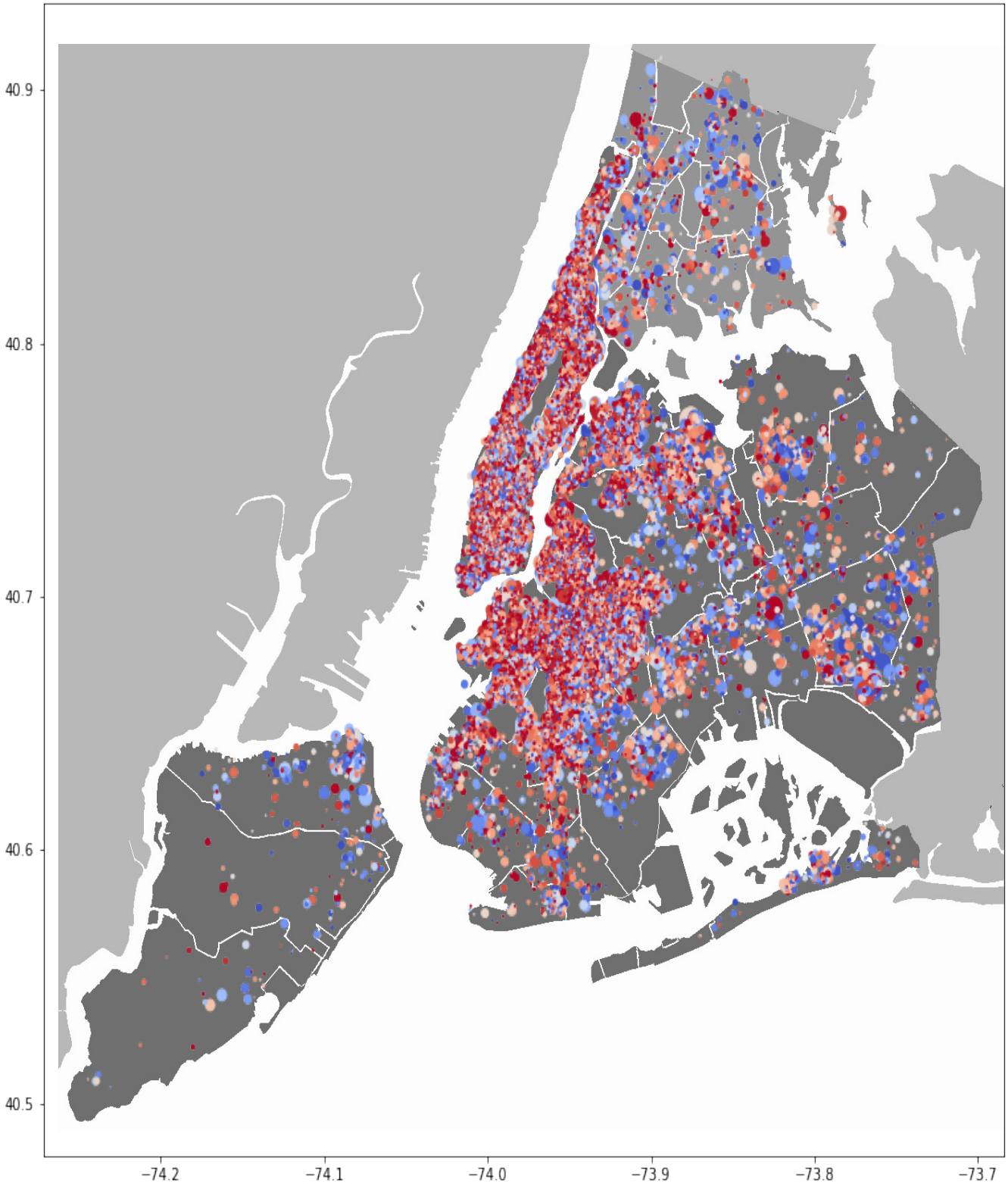
	Brooklyn	Manhattan	Queens	Staten Island	Bronx
5%	12,232,693.95	17,845,267.35	2,027,261.95	129,227.65	271,009.45
7%	17,125,771.53	24,983,374.29	2,838,166.73	180,918.71	379,413.23
10%	24,465,387.90	35,690,534.70	4,054,523.90	25,8455.30	542,018.90

Geographical Distribution of Airbnb

The geographical distribution of Airbnb is visualized on the New York City map. Each bubble represents an Airbnb listing. The color of the bubbles represents the number of days available for booking. The greater the number of days available, the warmer the color of the bubble. The size of the bubbles represents the total number of reviews for each listing. The more reviews there are, the larger the bubbles. The map is visualized in table 7.

The map shows that Manhattan and Brooklyn have the most Airbnb listings with a high density. The distribution of Airbnb listings in Queens is relatively sparse, and much fewer and sparser in Bronx as well as Staten Island. The warm color is mostly concentrated in Manhattan, Brooklyn, and the east of Queens, meaning that the Airbnb listings in these areas are available for more period of times. In Bronx, Staten Island, and the west of Queens, while bluer bubbles indicates that the listings in these areas are available for fewer days, the bigger size of these bubbles shows that these listings have a lot of reviews. That is to say, these listings attract many customers even with less availability. In Manhattan and Brooklyn, however, the red smaller bubbles indicate that a lot of listings attract fewer customers even though they are available for most of the year. The reason may be that the listings in Manhattan and Brooklyn have abundant supply but relatively limited demand, while the listings in Bronx, Staten Island, and the west of Queens have less supply, but relatively more demand given the cheaper prices of listings.

TABLE 7. THE GEOGRAPHICAL DISTRIBUTION OF AIRBNB IN THE NEW YORK CITY



Conclusion

To conclude, the New York City government is recommended to regulate the number of listings that each host should own and the number of days available for booking for each listing. The government should also collect rental tax revenue from Airbnb with an appropriate tax rate, based on the calculation of tax revenue with different tax rates and boroughs. Finally, the mapping of Airbnb, with the information on reviews and availability, should be publicized.

It was both surprising and alarming to find that as many as 5154 hosts own more than one listing and several hosts own abnormally large number of listings with a maximum of 327 listings. It was also worrying that listings on average are available for 175.83 days per year. The days available of listings are noticeably longer than a normal Airbnb listing, with 45.8% of listings available for more than half of the year. The government is suggested to investigate the reasons behind and pass necessary regulations. In terms of tax revenue, it is within our expectation that Manhattan is the most expensive among all the boroughs, while Manhattan and Brooklyn have the most Airbnb reviews. Thus, the government is expected to collect the largest amount of tax revenue from these two boroughs and the least revenue from Bronx.

It is important to acknowledge that, due to the limitations of the dataset, many questions remain unanswered. First and foremost, although we could infer from the data that Airbnb has impact on local housing availability, it is unclear how great that impact could be and to what extent it could affect the long-term housing prices. Second, since the dataset does not have information on total number of nights of staying for each listing, the function (price * number of reviews * minimum number of nights to stay * tax rate) had to be used to estimate tax revenue. This function could underestimate the amount of tax revenues from Airbnb, because some customers do not write reviews, and some stay more than minimum number of nights. All in all, the government is recommended to not overinterpret the results, which include but not limited to overstating the impact of Airbnb on local housing availability and prices as well as becoming overly confident in the estimated total tax revenue from Airbnb and the revenue from each borough.

If there is more time and budget to continue this study, I would recommend the future work to further analyze how the tax revenue from Airbnb varies with occupancy rate, minimum number of nights to stay, and other relevant variables. The future work could also focus more on the neighborhoods within the boroughs to find out the impact of Airbnb on each neighborhood, which should further help the government make effective relevant regulations.