How to Maximize Your Properties' Revenue:

Case Study on Airbnb in NYC

Contributed by Emma Sun

December, 2019

I. Introduction

1. Problem Statement

It is known that Airbnb is one of the biggest international home-sharing platforms that allows homeowners and renters ("hosts") to put their properties ("listings") online, so that guests can pay to stay in them. New York City (NYC) has one of the most active Airbnb markets in the United States with more than 50,000 properties flooding into the city since 2018. This report is aiming at examining the factors contributing to the income that Airbnb hosts can gain from each listing per month. Specifically, this project investigates the effects of 1) property location, 2) room type, 3) property's amenities on host's income of each listing per month and conducts machine learning process to predict the increase of host income. Accordingly, recommendations on property location, room types, property's amenities and other potential options can be provided for hosts to gain more income.

Obviously, the primary audience of this project is the existing and potential host of Airbnb. For hosts, participating in Airbnb is a convenient way to earn some incomes from their property. It is useful for them to figure out those factors more possibly contributing to higher incomes. Besides, one of the advantages of Airbnb is that it's up to each host to decide features and facilities in their properties. Each host would like to leverage such flexibility to determine an optimal property setting so that not only can hosts get satisfied incomes but their properties are popular among those guests also. Accordingly, my paper plans to help less-experienced hosts to approximately find that "sweet spot". Furthermore, given that Airbnb market in NYC is extremely comparative, potential hosts could refer to this paper for estimating their benefits gained from renting properties.

2. Structure of Report

This project comprises three sections and proceeds as follows. The first section will introduce the overall problem and general structure of this project. By the means of reviewing the previous research and problem background, this section will summarize imperative tools and approach for analyzing the overall problem, especially emphases the necessity of employing quantitative approach. Furthermore, this section will describe the data set used in this project in details and take six primary steps for cleaning up original data.

The second section will split the overall problem into five sub-tasks and report each sub-task's analysis outcomes in details. In the sub-task 1, the report will generally describe the statistical outcomes of critical variables and specially describe the popular characteristics of properties and hosts. This sub-task is regarded as a preparation for further analysis. In the sub-task 2, 3 and 4, this project will examine

the effect of listing locations, types of room and properties' amenities on host income or the average host income. Plenty of visualized outcomes will be provided to directly display the change of host income across different factors. In the sub-task 5, the project will leverage general linear model to interpret and predict host income. Model comparisons and reflections will be also provided.

The last section will summarize the final results based on the analysis above and discuss the potential feedbacks from audience, serious limitations of this analysis and potential directions of future research.

3. Motivation and Background

If you're planning on getting out of town, renting out your home while you're away can be a quick way to earn some extra spending money. However, earning money is not as easy as it seems through Airbnb's advertisement. As a potential data scientist, I am supposed to unveil the dynamics of market and figure relations between critical factors and host incomes. Hosts are more likely to share their houses only can they secure substantial incomes since welcoming strangers to stay in private houses is not always a delightful thing. If the income is limited, potential hosts would not like to bother to rent their properities. Therefore, this project is significant to hosts to determine whether enter into short-term renting market in NYC or not. Besides, as a rental ecosystem, Airbnb generates tons of data such as the density of listings across regions (cities and neighborhoods), price variations across listings, host-guest interactions in the form of reviews as well as prevalent listing characteristics. Given that this paper plans to analyze the change of income influenced by several factors, it is reasonable and imperative to conduct quantitative analysis to explore conclusive outcomes.

On the other hand, previous research primarily focused on the listing price prediction that shed light on this project when analyzing relations between host income and other factors. Some researchers came up with a best-performing model for predicting the Airbnb prices also based on a limited set of features including property specifications, owner information, and customer reviews on the listings. They also include some hidden factors and interactive terms, such as the personal characteristics of the owners (Kalehbasti, et al., 2019). LI Yang proposed the method of Multi-Scale Affinity Propagation (MSAP) aggregating the house appropriately by the landmark and the facility. In each cluster, they applied Linear Regression model with Normal Noise (LRNN) predicts the reasonable price, which is verified by the increasing number of the renting reviews (LI, et al., 2016). Those two reports as well as other Op-Eds provided insightful ideas.

4. Dataset

The original data used in this project was downloaded from the website of Inside Airbnb, named as *listings*. This is a comprehensive data set containing a total 50,599 listings with 106 variables. The 106 variables encompass names of houses and hosts, location information (latitude, longitude, boroughs and neighborhoods), room types, nightly price, minimum number of nights for staying, reviews amount, and so forth. The data are relatively complex since they encompass multilevel and multifaced information of hosts, guests and properties. In terms of the copyright status, this dataset is a public dataset, provided by an anti-Airbnb lobby group named Inside Airbnb. This group updates listing data every month to follow any change in market. Besides, no obvious privacy violation will be made based on my analysis so far. Given that the goal of this project is to explore how to increase host's income, this paper specially focused on those highly related and valuable features and re-organized the data set for analysis as follows:

4.1 Explore Potential Variables

First of all, this project chose 33 variables directly related to host-guest interactions and characteristics of listings based on common sense and empirical experience. The chosen variables were classified to three groups: host information, listing characteristics and time variables. The time variables were only used to determine the time scope of data. Table 1 shows the description of those variables.

Table 1. Variables Description

Host Information	
id	Host's unique ID
host_name	Host's name
host_response_rate	Rates of host responding to guests
host_listings_count	Number of listings hosts own
host_is_superhost	Describes highly rated and reliable hosts (true/false)
host_has_profile_pic	Whether or not hosts has pictures on their profile (true/false)
host_identity_verified	Whether or not the the identity of the host has been verified (true/false)
instant_bookable	This is a feature that hosts can select for their listing which removes the approval process. Instead of being notified that someone is interested in your place and then giving them the thumbs up/down – guests can just choose their dates, book, and discuss check-in plans later (true/false)
require_guest_profile_picture	Whether require guest upload pictures (true/false)
require_guest_phone_verification	Whether require guest verify mobile phone (true/false)
Listing Characteristics	
room_type	Room types: Entire home/apt, Private room, Shared room and Hotel room
property_type	Type of property
bed_type	Type pf bed. eg: real bed, pull-out sofa, futon, airbed, and couch
beds	Number of beds included in the property
minimum_nights	Minimum number of nights a guest can stay
maximum_night	Maximum number of nights a guest can stay
number_of_reviews	Number of reviews that previous guests have left
security_deposit	Required deposit
cleaning_fee	Cleaning costs

square_feet	Amount of space
guests_included	Number of guests the booker wants to include
extra_people	Cost of additional person per night
latitude	Locational coordinate
longitude	Locational coordinate
accommodates	Number of guests the rental can accommodate
bedrooms	Number of bedrooms included in the rental
bathrooms	Number of bathrooms included in the rental
number_of_reviews	Number of reviews that previous guests have left
neighbourhood_group_cleansed	Neighborhood groups in NYC
amenities	Amentities in property
review_scores_rating	review scores
Time Variables	
calendar_updated	Date listing updated
last_review	Data of most recent review

4.2 Remain Listings after 2018

Secondly, the original data contains listing information more than 90 months ago. This project determined to remain listings posted after 2018 since listings information updated before 2018 seems too out-of-date to reflect the current situation of listings in NYC accurately. This project employed two approaches to discard data before 2018. The first one kept the values in the variable "last_review" after December 31, 2017 (“2017-12-31”), which reduced the number of listings to 20,170. The assumption of this attempt was that if the most recent review was received before 2018, it implied that the listing was no longer popular or the host would not like to rent their properties. The other one kept the values in the variable "calendar_updated" after “24 months ago” (including 24), which reduced the number of listings to 42,623. Obviously, the first approach was too conservative to show the listings in the NYC after 2018 since it is common that guests would not like or forgot to upload reviews. Therefore, this project determined to employ the second approach.

4.3 Address Outliers

Thirdly, this paper examined the missing values in the data set and 13 variables in total contained missing values which were addressed as follows:

1) More than 99% of values in the “square_feet” were missing values, thus, this variable was deleted.

2) Less than 1% of values in the “beds”, “bedrooms” and “bathrooms” were missing values. Instead of deleting those value, this project replaced missing values with their modal values respectively.

3) The type of “security_deposit” or “cleaning_fee” was object, thus missing values in both variables were replaced with value '$0'.

4) The missing values in the variables 'host_reponse_rate' and "review_scores_rating' were replace with value zero.

5) Those rows containing missing values in the “host_identity_verified”, “host_name”, “host_is_superhost”, “host_has_profile_pic”, and “host_listings_count” were directly deleted.

So far, the data set contained a total 42,143 listings with 30 variables.

4.4 Clear Messy Characters in Values

Besides, in order to better process the analysis, this project removed dollar mark ($) from values in the “price”, “security deposit”, “cleaning fee”, and “extra_people”, and thus, their types were transformed from object to float or integer. The type of “host_response_rate” was also transformed from object to float. Additionally, messy characters in the “amenities” were also deleted as indicated in the codes.

4.5 Create Response Variable

The next critical step was to create the response variable “income” to indicate the host income of each listing per month (“host income” for short as follows) for this analysis. Since over half of hosts had one more listing and some of hosts were companies not individuals, it is more reasonable to analyze host income for each listing rather than each host’s incomes in total. The response variable was extrapolated with the formula as follows:

$$\textit{Total Income} = \textit{Income of Listing} * \textit{Number of Booked Nights} / 12$$

$$\textit{Income of Listing} = \textit{Price of Listing} + (1\text{-}3\%) * \textit{Cleaning Fees}^{1}$$

$$\textit{Number of Booked Nights} = \textit{Number of Reviews} * \textit{Minimum Nights}$$

So far, the data set contained 42,143 listings with 31 variables. Among the 31 variables, there were 19 numerical variables and 12 categorical variables.

4.6 Address Outliers

Last but not the least, this project addressed outliers in the 18 numerical variables through discarding extreme values. The first approach was to discard values above the upper limits (Q3 + 1.5 * IQR, Q3: upper quartile, IQR: interquartile range) in the boxplots respectively. Meanwhile, zero values in "host_listings_count" and "price" were also deleted. This step reduced the data set to 534 listings. This

[1] Airbnb charges 3% host service fee that is calculated from the booking subtotal (the nightly rate plus cleaning fee and additional guest fee, if applicable).

approach was obviously too conservative. The other approach only addressed those obvious extreme values and manually set maximum values and minimum values of “income”, “host_listings_count”, “maximum_nights”, “minimum_nights”, and “price” based on those boxplots. This project restricted the income within 200,000; the number of host’s listings within 1,000; the number of maximum nights within in 1,125; the number of minimum nights within 500; and the price of listings within 5,000. This attempt reduced the number of listings to 32,739.

Through the six steps above, the data set named “key_listings” was fully prepared for the analysis as follows.

II. Results

1. Sub-task 1: Summary Statistics

1.1 Description of Task

The first sub-task was to make statistical descriptions on critical variables. This subsection conducted series of descriptive statistics and data evaluation to get better understanding of this dataset. Based on the findings in this subsection, some critical variables could be picked up to examine their influence on host incomes. Some critical findings were shown as follows.

1.2. Presentation of Result

1.2.1 Descriptive Statistics of Numerical Variables

The average host income of each listing per month was approximately $2191.76. As parts of the income, the average cleaning fee was $56.47, while the mean price was $139.4. On average, guests were asked to book listings at least a week. Besides, the average responding rate of hosts was 74%, indicating host could response to most guests. Table 2 displayed the descriptive statistics of numerical variables in details.

Table 2. Descriptive Statistics

Variables	N	Mean	*SD*	Min	Max
income	32739	2191.76	7347.33	1.67	194617.5
accommodates	32739	2.94	1.94	1	16
bathrooms	32739	1.15	0.43	0	15.5
bedrooms	32739	1.19	0.75	0	21
beds	32739	1.59	1.19	0	40
cleaning_fee	32739	56.47	55.04	0	1000
extra_people	32739	17.06	25.62	0	300
guests_included	32739	1.62	1.25	1	16
host_listings_count	32739	8.12	46.55	1	975

host_response_rate	32739	0.74	0.41	0	1
latitude	32739	40.73	0.06	40.51	40.92
longitude	32739	-73.95	0.05	-74.25	-73.71
maximum_nights	32739	560.09	530.39	1	1125
minimum_nights	32739	6.28	14.74	1	456
number_of_reviews	32739	36.12	54.38	1	675
price	32739	139.4	148.86	10	5000
review_scores_rating	32739	92.33	14.76	0	100
security_deposit	32739	180.1	414.9	0	5100

1.2.2 Popular Characteristics of Properties

Properties have some popular characteristics. First of all, among the 129 various amenities in the properties in NYC, the top 10 amenities were “Wifi” (98.40%), “Heating” (95.69%), “Essentials” (95.47%), “TV” (92.96%), “Kitchen” (90.27%), “Smoke detector” (89.82%), “Air conditioning” (87.59%), “Hangers” (81.36%), “Carbon monoxide detector” (74.83%), and “Hair dryer” (74.21%). Secondly, Manhattan and Brooklyn were the most popular Airbnb neighborhood groups in NYC. 41.79% of listings were located in Manhattan while 41.12% of listings were located in Brooklyn. Thirdly, among the 35 types of properties, more than three quarters of properties (75.61%) were “Apartment”. There were also other popular types of properties, such as “House” (9.76%), “Townhouse” (4.22%), “Condominium” (3.33%) and “Loft” (3.32%). There are four types of room: entire home or apartment, private room, shared room and hotel room”. More than half of rooms (51.52%) were entire home or apartment, while 45.23% of properties contained private rooms. As for the five types of beds, 98.59% were “Real Bed”. The other four types were “Futon” (0.55%), “Pull-out Sofa” (0.5%), “Airbed” (0.24%) and “Couch” (0.12%). In addition, 59.7% of properties can be reserved instantly, while 40.3% of properties cannot.

1.2.3 Popular Characteristics of Hosts

There are also some interesting findings across the hosts. Two mysterious hosts both named Christian owned the most listings that were 975 listings. The most popular host was Dona from Queens who received the most reviews up to 675 reviews. Almost all the hosts (99.8%) uploaded their profile pictures. Approximate all guests were required to verify their phones (97.12%) and to upload their profile photos (97.39%). However, only half of hosts’ identities (55.57%) were verified and around three quarters of hosts (75.15%) were super-host.

2. Sub-task 2: Host Income and Listing Location

2.1 Description of Task

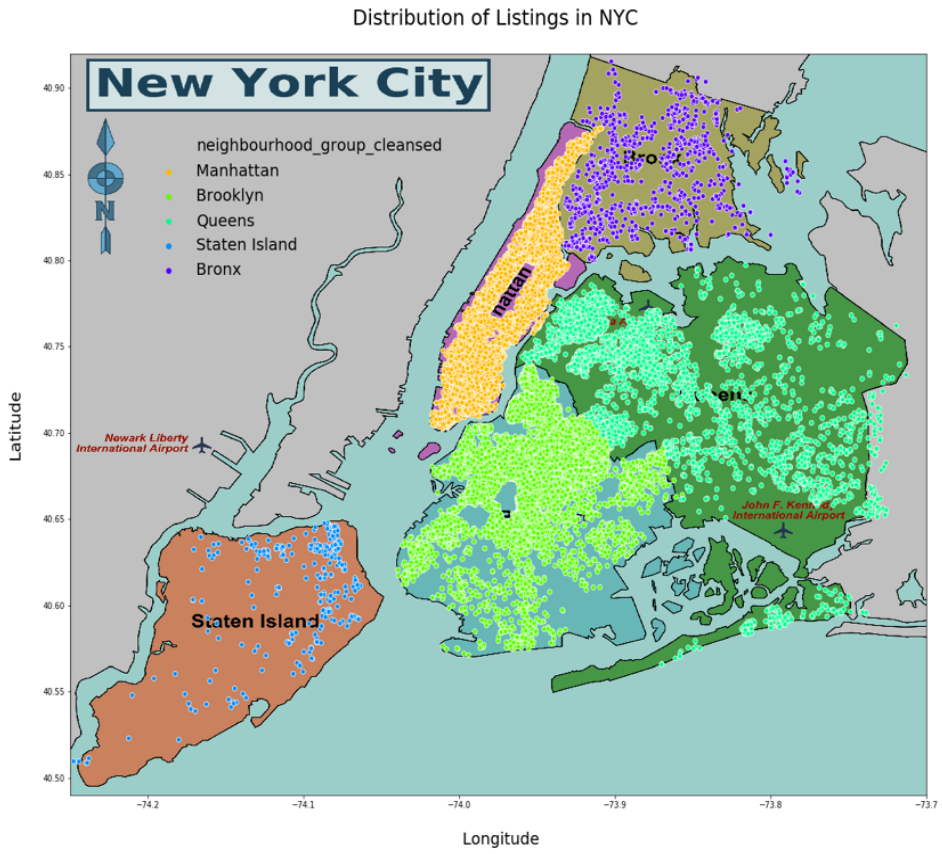
This subsection was aiming at examining the correlation between host income and listing locations. First of all, this section displayed the geographical distribution of listings in NYC. Secondly, this section examined the relation of the average host income and neighborhood groups. Furthermore, this section investigated and visualized how host income varied across the neighborhood groups.

2.2. Presentation of Result

2.2.1 Listings Distribution in NYC

Graph 1 vividly indicted the distribution of listings in NYC. Spots in Manhattan and Brooklyn were much denser compared to other neighborhood groups, indicating Manhattan and Brooklyn were the two most popular neighborhood groups among hosts.

Graph 1. Distribution of Listings in NYC

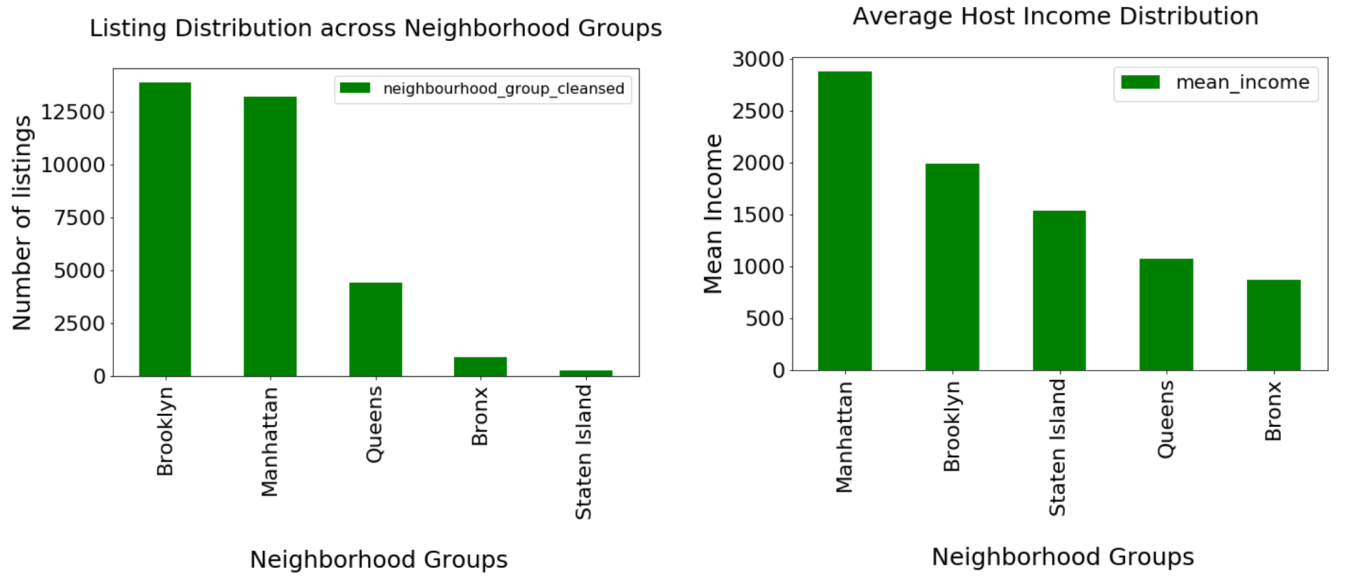


2.2.2 Average Host Income across Neighborhood Groups

Graph 2 indicated that the number of listings and average host income in different neighborhood groups. Despite the amounts of listings in Brooklyn and Manhattan being similar, average host income in Manhattan was significantly higher than the average host income in Brooklyn. The mean host income in Manhattan is $2877.75, while the mean host income in Brooklyn is $1994.7. This finding is reasonable

since per capita income in Manhattan is always the highest among the all neighborhood groups for a long time. Additionally, although less than 1% of listings were in Staten Island, the average host income in Staten Island was much higher than it in Queens and Bronx. Potential hosts seemingly should explore the market in Staten Island.

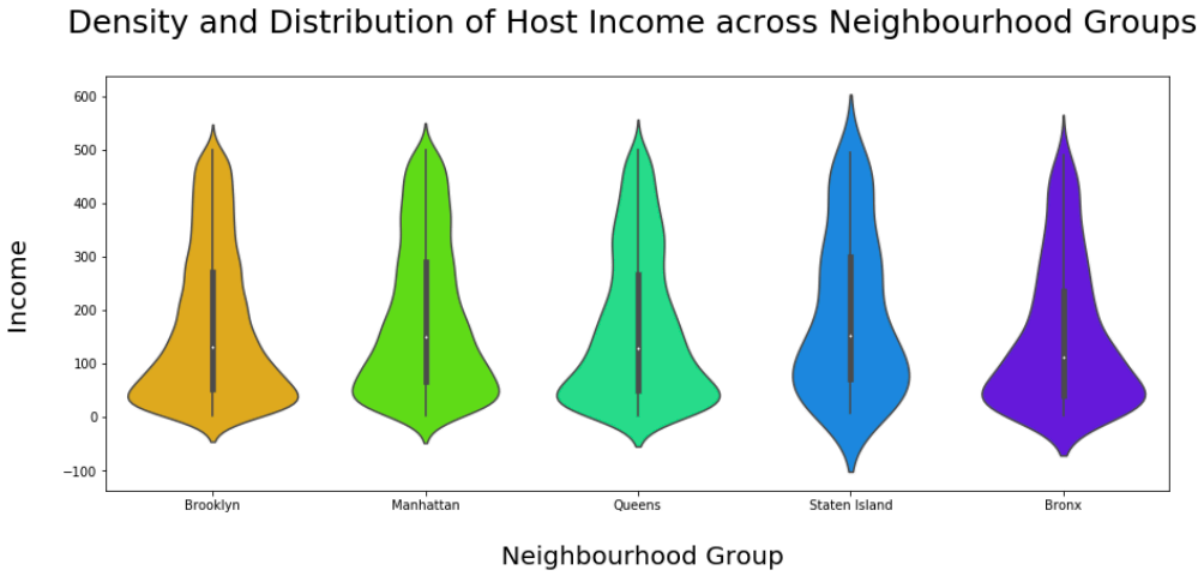
Graph 2. Distribution of Listings and Average Income across Neighborhood Groups



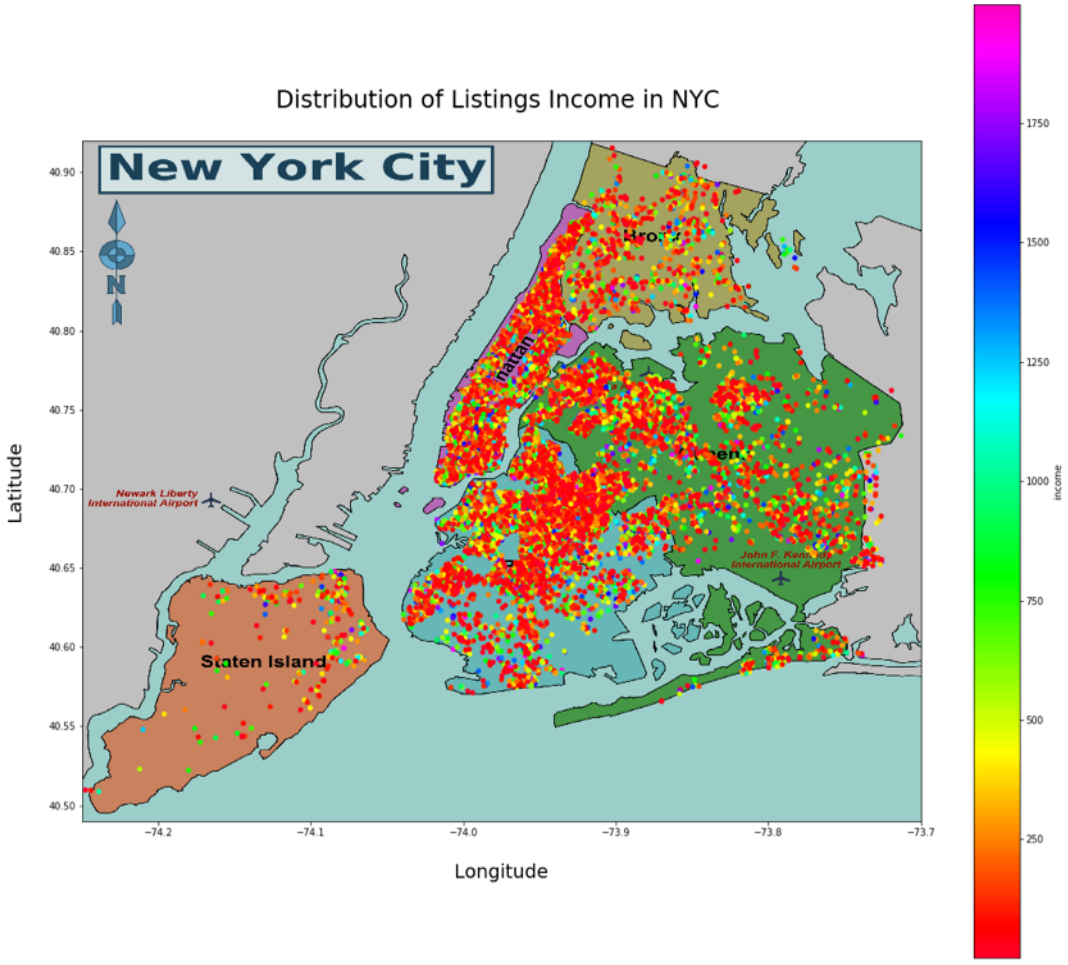
2.2.3 Host Income across Neighborhood Groups

Violin plots in Graph 3 indicated that host income did not uniformly distribute across neighborhood groups. Compared to the other four neighborhood groups, host income in Staten Island distributed more evenly. The color of scatters in the region of Staten Island was heterogeneous shown in Graph 4, indicating such relatively uniform distribution. Most host earned income less than $500 in Manhattan and Brooklyn indicated by the dense red scatters in Graph 4. Based on the Graph 3 and 4, the average host income, to some degree was unable to indicate the actual host income in different regions.

Graph 3. Density and Distribution of Host Income across Neighborhood Groups



Graph 4. Distribution of Listings Income in NYC



Additionally, based on correction coefficients, the correlations between host income and different listing location was very weak (all coefficients were less than 0.1, see results in code Section IV, 3, 3)), thus location was not be regarded as one of critical factors in regression analysis.

3. Sub-task 3: Host Income and Room Types

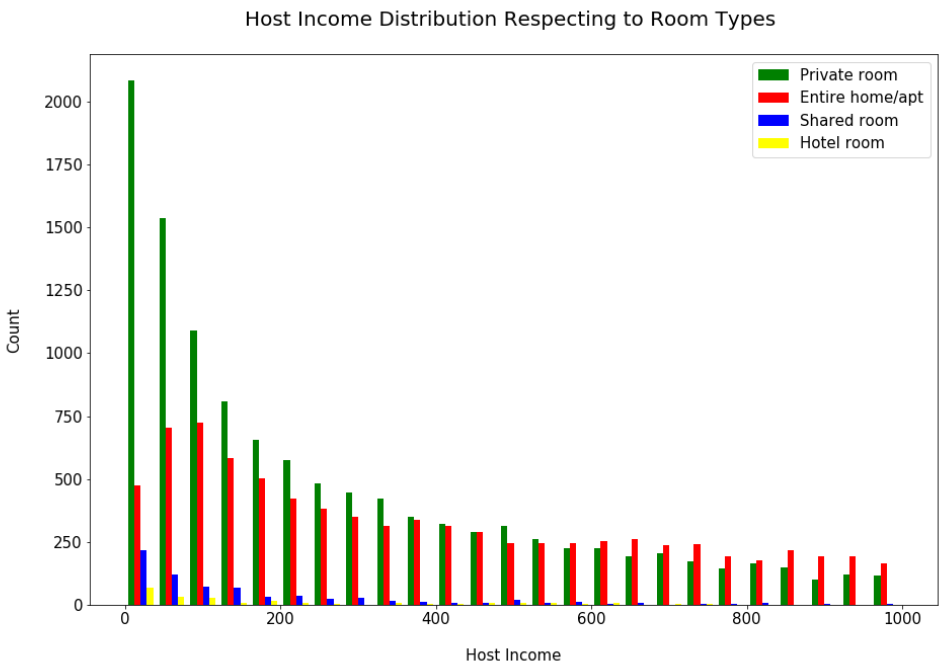
3.1 Description of Task

This subsection was aiming at examining the correlation between host income and room types, and answered the question: How did host income vary across the four room types?

3.2. Presentation of Result

Graph 5 indicated that the distribution of host income across different room types. Majority of host incomes of listings with private rooms were in the range from $0 to $100. With the increase of host income, the number of listings with private room sharply decreased. Compared to other room types, the number of listings with private room was larger at every level of host income when host income was less than $500. On the other hand, compared to listings with private room, host income of listings with entire home or apartment distributed more evenly. When host income was larger than $500, the number of listings with entire home or apartment was larger than the number of listings with private room at every level of host income. Besides, based on correction coefficients, the correlations between host income and listings with entire home or apartment as well as listings with private room were relatively high (see results in code Section V, 2), thus this project took it into consideration in regression analysis.

Graph 5. Host Income Distribution Respecting to Room Types



4. Sub-task 4: Host Income and Amenities

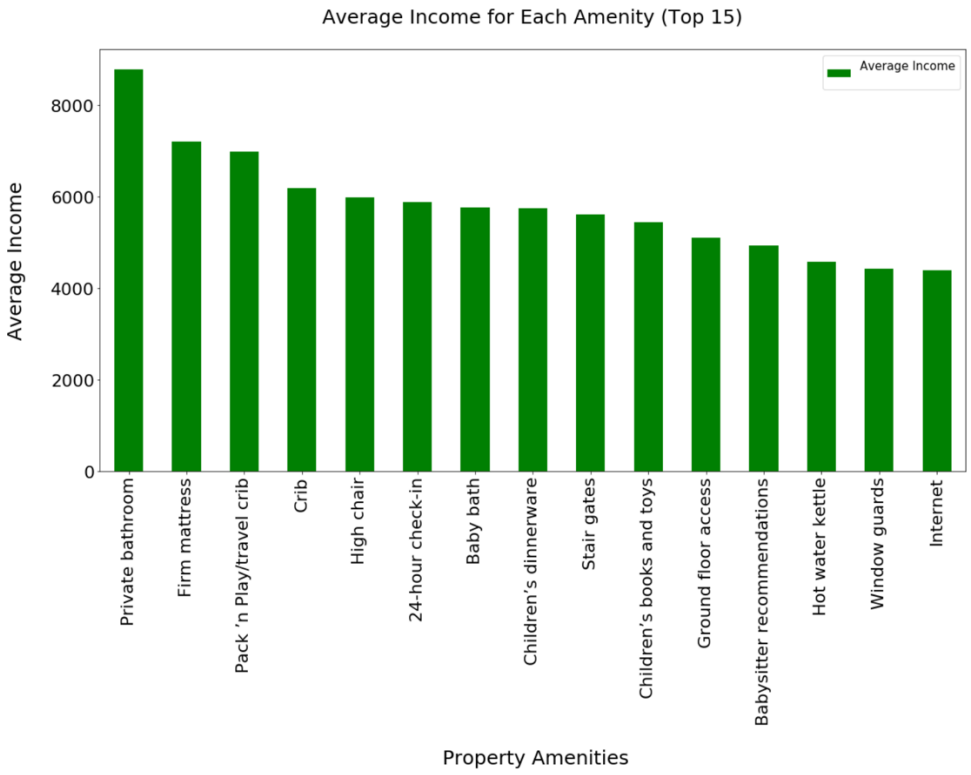
4.1 Description of Task

This subsection was aiming at examining the relations between host income and property amenities, and answered the question: How did average host income vary across the properties' amenities?

4.2. Presentation of Result

Graph 6 indicated the top 15 average host income of listing with different property amenities. The average host income of listings with private bathroom was largest. Among the 15 amenities, 9 amenities were child-friendly features and facilities that were cribs, high chair, baby bath, children's dinnerware, stair gates, children's books and toys, babysitter recommendations, and window guards. Seemingly, child-friendly facilities are important for securing higher incomes.

Graph 6. Average Income for Each Amenity (Top 15)



Based on the outcomes of T-test, the mean host income of child-friendly listings was $5674.56, while the mean host income of non-child-friendly listings was $2579.43. This remarkable difference was significant (t=11.234, p=0.000). The mean host income of listings with private bathroom is $8774.51, while the mean host income of listings without private bathroom is $2908.49. This remarkable difference was also significant (t=11.234, p=0.000). According, it is necessary that hosts should take child-friendly

amenities and private bathrooms into consideration when they determined to rent their properties, and highlight those amenities to draw potential guests' attention.

However, it is worth noticing that less than 1,000 listings were equipped with children facilities. In other words, such higher income might indicate some special cates rather than represent the general circumstances. Nevertheless, in spite of being imperative to consider the frequency of each amenities, this finding was still meaningful to hosts.

5. Sub-task 5: Regression Analysis

5.1 Description of Task

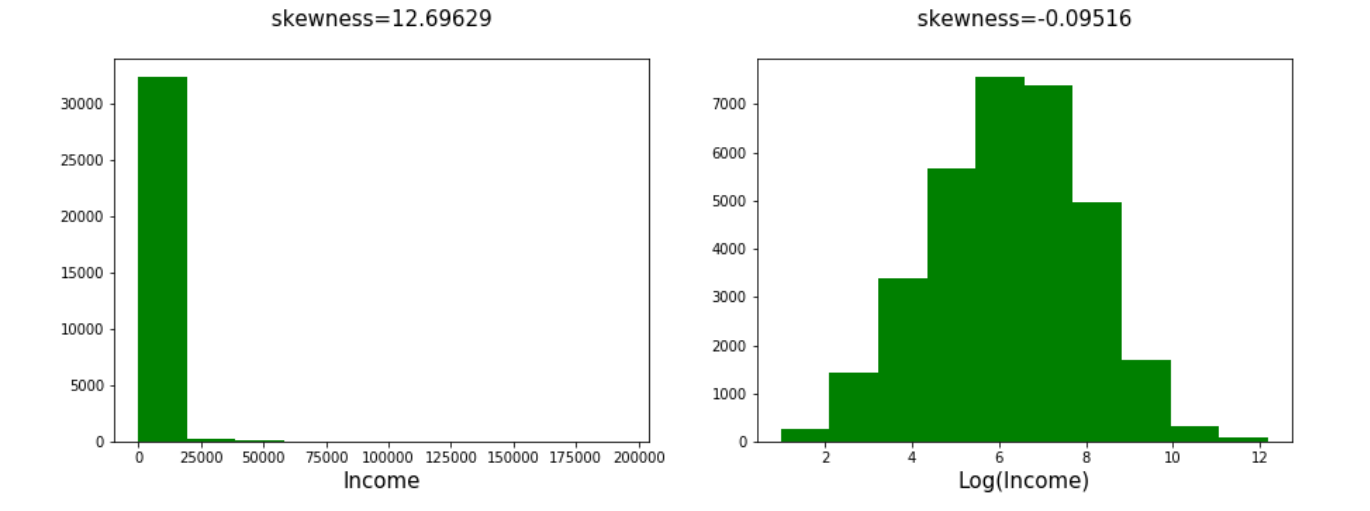
This subsection was aiming at examine the factors contributing to host income through creating two general linear models. Before processing the analysis, this subsection prepared the dataset by logarithmically transforming the response variable "income", assigning values for variable "amenities" and encoding those categorical variables. The second half subsection conducted regression analysis on factors contributing to the increase of host income.

5.2. Presentation of Result

5.2.1 Prepare Dataset for Regression Analysis

First of all, this subsection transformed the response variable "income". Based on the histograms of distribution of "income" and "log (income)" (see Graph 7), this project logarithmically transformed "income" to "log_income" that was used in regression analysis.

Graph 7. Distribution of Income and Log Transformed Income



Secondly, there are numerous amenities in each listing. It was unnecessary and impossible to employ the built-in encoding function in *pandas* to separate all different amenities combinations. Instead, this project determined to assign values to "amenities" for regression analysis as follows:

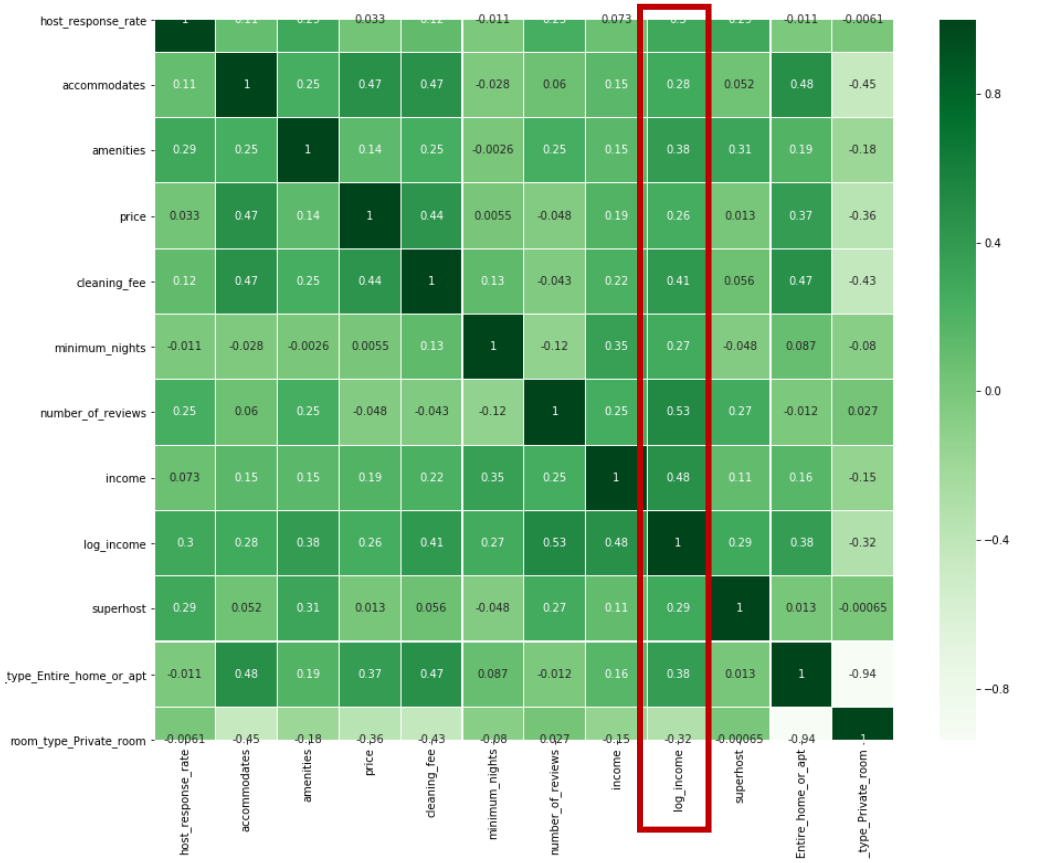
Each sub-amenity $i \in$ {Wifi, Heating, Essentials, ……} was assigned the weight $w(i)$ based on the formula: $w(i) = f(i) / f_{max}$ where $f(i)$ is the frequency of i in the data set and f_{max} is the maximum frequency of certain amenity, that was the frequency of Wifi. Theoretically, the more frequent a specific amenity was, the greater its weight was. The sum of the weights s is used to represent the value of each set of amenities: $s = \sum w(i)$. In this way, each listing had its correspondent value of amenities, and listings equipped with more common amenities would have larger values.

Furthermore, other categorical variables were also encoded in this subsection. One the one hand, this project classified the types of property as apartment and non-apartment since more than three quarters of properties were apartments. All the Boolean variables were encoded by assigning "1" to "t" and "0" to "f". On the other hand, the room types were spited into four dummy variables. Therefore, 29 variables with 32,739 listings were used in the regression analysis.

5.2.2 Model Creation and Interpretation

After examining the correlation coefficients between variables with the response variables, this subsection created a simple linear model containing 10 independent variables whose correlation coefficients were more than 0.25, named Model 1. Correlation levels of those 10 variables were shown in Graph 8.

Graph 8. Correlation Heatmap



According to R square, 62.71% of variances could be interpreted by this model. All variables were significantly correlated to the response variables. Some critical findings were shown as follows (see Table 3):

1) With any additional increase of host's responding rate, the host income was predicted to significantly increase $1.5342.

2) The increase of the number of accommodates was negatively correlated to the host income.

3) The super host's income was predicted to have $1.40 higher than non-super host's income, holding all other independent variables constant.

4) The host income of listings with private rooms were predicted to have $1.76 higher than host income of listings without private rooms, holding all other independent variables constant.

5) The entire house or apartment was predicted to have a $3.59-higher host income compared to other room types, while the private room was predicted to have a $1.76-higher host income compared to other room types, holding all other independent variables constant.

[2] The value was anti-logarithmic income.

In order to examine the interaction effect of being a super host on the rates of hosts' responses to guests, this project added an interaction item "superhost* host_response_rate" to Model 1, named Model 2. According to the R-squared and RMSE (see Table 3), the fit of Model 2 slightly improved. The effect of being super host on income became more obvious compared to Model 1. The super host's income was predicted to have $2.31 higher than host's income who was not super host, holding all other independent variables constant. However, such advantage was moderated by host's responding rate; with the increase of responding rate, the advantage of being super host decreased.

Based on the findings above, this project suggests that for the sake of higher income, hosts should response to guests' requests as much as possible. Through becoming super hosts and renting entire home or apartment, hosts might increase their income slightly.

Table 3. Model Summaries of No Growth Model and Linear Growth Models

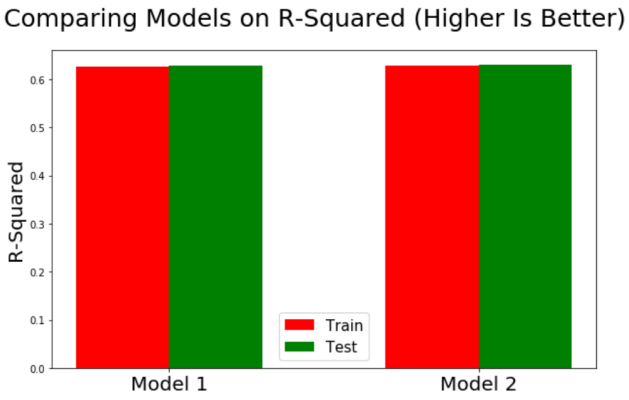
	Model 2			Model 1		
	coef	*std err*		*coef*	*std err*	
Intercept	2.9665	0.04	***	3.0037	0.04	***
room_type_Private_room	0.5724	0.034	***	0.5657	0.034	***
minimum_nights	0.0346	0	***	0.0347	0	***
accommodates	-0.0383	0.004	***	-0.038	0.004	***
superhost	0.838	0.052	***	0.3361	0.015	***
host_response_rate	0.4763	0.017	***	0.4279	0.016	***
superhost*host_response_rate	-0.5512	0.055	***	—	—	
number_of_reviews	0.0165	0	***	0.0165	0	***
cleaning_fee	0.0073	0	***	0.0073	0	***
amenities	0.0465	0.002	***	0.0466	0.002	***
room_type_Entire_home_or_apt	1.2878	0.035	***	1.2772	0.035	***
price	0.0012	4.78E-05	***	0.0012	4.79E-05	***
N	32739			32739		
Adjusted R^2	0.6283			0.6271		
RMSE	24489787			24990448		

* p< .05, ** p< .01, *** p< .001.
RMSE: Root Mean Square Error (RMSE)

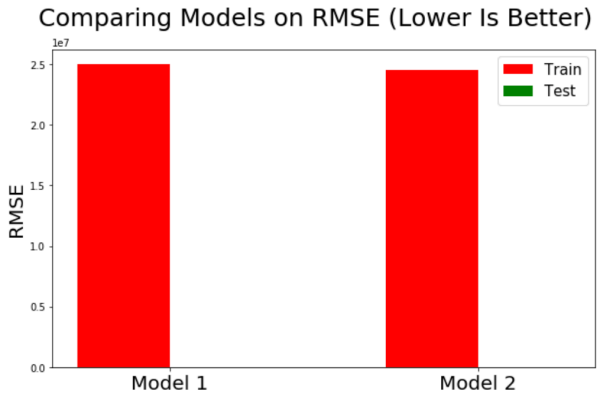
5.2.3 Out-of-sample Testing

Furthermore, this project tried to find out whether the models above would better or worse at forecasting in the future on data that was not part of the original training process. Therefore, the dataset was randomly divided into two sections where 80% of the original data were at a training data set while 20% of the original data were at a testing data set. RMSE from both Model 1 and Model 2 in testing data set was extremely small, indicating both of models could predict the future host income well.

Graph 9. R-Squared and RMSE Comparisons



Model 1 (Train): 0.627
Model 1 (Test): 0.629
Model 2 (Train): 0.628
Model 2(Test): 0.630



Model 1 (Train): 24990448
Model 1 (Test): 1.090
Model 2 (Train): 24489787
Model 2(Test): 1.089

III. Conclusion

1. Final Result

This project explored and modeled Airbnb listings data in NYC in 2018 and 2019 in order to help hosts secure higher incomes of each listings per month. Based on the analysis, the average host income of each listing per month was approximately $2191.76. Such income depended on price of listings, minimum bookable nights, cleaning fees and service fees charged by Airbnb. As expected, the most frequently listed amenities were Wi-Fi, Heating and Essentials. However, to secure higher incomes, hosts should create a child-friendly environment by setting facilities such as crib, high chair, children's books, toys, dinnerware as well as safeguard facilities such as stair gate and window guards. As for the room types, for the sake of a long-term and higher income, listings being entire house or apartment performed better than those being private room. Private rooms were more popular among guests having short-term demands, while the entire houses or apartment were more popular among guests with a long-term living demand. Given that host income would increase with the increase of minimum bookable nights, entire house or apartment would contribute to a higher income from the perspective of long term.

When it comes to the location of listings, more than 80% of listings were located at Brooklyn neighborhood group and Manhattan neighborhood group and the mean host income in Manhattan was statistically $1,000 higher than mean host income in Brooklyn. However, this project would not suggest hosts rent their properties at comparative Manhattan. Instead, hosts should explore the market in Staten Island where people can get away from the highly stressful life in NYC and enjoy the natural sceneries.

So far, the number of listings at Staten Island was the smallest among the five groups, but the mean host income at Staten Island was much higher than the mean host income at Queens and Bronx. Therefore, hosts at Staten Island should leverage their advantages of location.

The simple machine learning model of host income based on month based on 10 variables indicated that the linear regression model explained the host income variation in both training data set and testing data set well (as measured by the coefficient variation RMSE). The price and cleaning fee of listings, number of reviews, and the number of minimum bookable nights were all positively corelated to the increase of host income, while with the increase of the number of accommodates, host income was predicted to decrease. Besides, being super hosts and higher rates of responding to guests' demands also contributed to a higher mean host income. However, the effect of being super host on the host income was moderated by the increase of responding rate. At the initial stage, being super host contributed to a higher income. However, with the increase of the responding rates, the degree of such contribution decreased gradually. Therefore, regardless of being super host or not, to secure a higher income, hosts should pay more attention to their guests' demands and respond to messages as soon as possible.

To sum up, for the sake of the higher income, timely response, guest' privacy (indicated by private bathroom), spacious room (indicated by entire home) and child-friendly facility are the four most significant features that hosts should take into a serious consideration when renting their properities.

2. Discussion

Generally, the findings from this project were reasonable and provided a relatively comprehensive picture of listings and hosts in NYC. After reading this report, the audience are expected to have a clear understanding on how to optimize their listing characteristics and personal service to secure a higher income. The section addressing the unclean data and processing machine learning were quite interesting to data scientists but a little bit tedious to the audience. Therefore, the potential audience can directly skip to the sub-sections of sub-task 2, 3, and 4 where several graphs vividly displayed the relations between host income and critical features. The audience should also pay more attention to the section of conclusion to explore the outcomes of this project. From personal perspective, the most interesting part in this project is the Sub-task 4 that examined the average host income among different amenities. It is unexpected to find that children facilities make a remarkable influence on host income, but on second thought, this finding is reasonable. Listings equipped with children facilities are more likely to bring satisfied living experience and good feedback from guests, hence hosts are more likely to gain good reputation and deserve higher incomes.

However, this project has some inherent limitations. First of all, due to the limitations of the variables in the original data, the formula of host income has its weakness. The estimate of bookings amount is rather conservative and unreliable, since it is likely that properties have more bookings than number of reviews or guests are also likely to stay longer than the required minimum nights. On the other hand, any cost was not included into this formula, thus, the reliability of this project was violented. For instance, it is not safe enough to draw the conclusion that a child-friendly property must contribute to a higher income since the hosts need to pay for the maintaining fee for the facilities in their properties but we have no idea about the amount of that expense.

Secondly, when processing the regression analysis, this project only examined the interaction effect of being super host on responding rates. However, other latent interaction effects still remained unaddressed, which led current models to the danger of multicollinearity. On the other hand, despite other correlations, due to the limited technical capacity and statistics knowledge, this project only employed the linear regression model to interpret variables, which perhaps further decreases the reliability of outcomes in this project.

Respecting to conducting further analysis on Airbnb in NYC, the change of host income across time should be taken into consideration by combining other data sets with existing data. Besides, it is also interesting to explore what are the statistically significant distinctions between super hosts and ordinary hosts. As for the machine learning process, other variables should be included into model to reduce the RMSE error. Meanwhile, other models such as random tree models should be employed to compare which model will better predict the host income. I hope I can accurately figure out which factor has the strongest impact on host income, so that, the scale of recommendations on the listing setting will be narrowed down and hosts will customize their listings for predictable profits. Furthermore, I also would like to change my audience to city planners and to analyze how to arrange the distribution, density and inside setting of listings respecting to generating more fiscal revenues in NYC.

References

Kalehbasti, P., Nikolenko, L., Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis.

LI Y., PAN Q., YANG T., GUO L. (2016). Reasonable price recommendation on Airbnb using Multi-Scale clustering. Chinese Control Conference (CCC). 2016,7038-7041. doi: 10.1109/ChiCC.2016.7554467