Multi-dimensional Analysis On New York City's Airbnb Data

Jiwon Choi
Department of Computer Science
Northwestern University
Evanston, IL 60208, USA
jiwonchoi2021@u.northwestern.edu

***Abstract*—In order to provide more insights to users, Airbnb data on New York City have been utilized to perform a multi-dimensional analysis. The analysis contains general exploratory data analysis, statistical and machine learning models to provide the best intuition. The EDA shows the overall distribution of data, and basic relations between them. The statistical model provides more profound analysis on the relations between variables. Although multiple hypothesis have been made, a lot of tests resulted to be failed to reject the null hypothesis. Multiple machine learning models have proposed to predict the superhost and the price, but the price estimation model reported a failure and needs a further improvement.**

I. Introduction

Data is a collection of information that can be stored by using any type of media. It can be both useful, and useless, so yielding a meaningful information through the data science pipeline is indispensable. The data science pipeline contains multiple steps: (1) data collection, (2) data cleaning and management, (3) exploratory data analysis, (4) data modeling, and (5) data interpretation. In this paper, we will be using a data collected by Airbnb, especially the New York City's data. This data is composed with a numerica and the text type of data is also available.

The era of post-COVID is coming, and many people started to plan their trip ahead. In order to provide a meaningful data to both suppliers (Airbnb hosts) and demanders (Airbnb users), this multi-dimensional analysis on Airbnb data is needed. What users may take a look at before making a final decision is not only the price and but also the reviews that other users have written before, and the description provided by the host. In case of hosts, they need to manually search what other hosts' are doing in order to competitively compare themselves to one another. With those needs, some type of more intuitive, concise, and informative data visualization is needed, which can be easily understood and less time-consuming.

In order to tackle this problem, further analysis on the followings are needed:

- General analysis on the properties, reviews, and price of the place.
- Given the feature, provide a reasonable price and superhost estimation for listings.
- Word cloud plots so that users and hosts would know the summarized keywords.
- Sentiment analysis per each county in New York City, to suggest users which area to consider.

II. Data Collection

In order to analyze the proposed problem, the data set provided by Airbnb will be used. They provided the data here (http://insideairbnb.com/get-the-data.html), and they have a subsection for New York City which will be used in this study. Under New York City, there are multiple data set that Airbnb has provided. Among 7 data set that they have, this study utilized 2 of them: `listings_detail.csv`, and `reviews_detail.csv`. The data that Airbnb has provided is in a `*.csv` (comma-separated values) format, one of semi-structured formats. Compared to other two formats like structured and unstructured ones, this semi-structured format is one of the widely used one among data scientists. Note that description of each field (metadata) can be found here (https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=982310896).

The Airbnb data set has both traditional and opinion data available. For traditional type of data, it has an information including host, their url, reviewer's id, and etc., which is about the user information that is registered or stored to Airbnb's platform. In case of opinion data, the collected data contains users' reviews about the places that they have stayed. and this data will later be utilized to perform some text-related analysis including word cloud plot.

III. Data Cleaning & Management

Data cleaning step is an important and essential step in data science. It is a time-consuming and iterative process, yet easily overlooked by many scientists. Before actually getting into the analysis step, adequately cleaned data set is required. Erroneous data may result a misleading conclusion, and the higher the quality, the more the accurate result, and ultimately yields the better decision. Yet, overcleaned data will oversimplify the original data set, and lead to a misrepresentation. In this study, the following cleaned data will further be examined via exploratory data analysis (EDA) in the later section to validate whether the data has cleaned appropriately or not.

This study was done for CS 396: Introduction to the Data Science Pipeline at Northwestern University, under the guidance of Dr. Huiling Hu.

There are two types of data in the social science definition: primary and secondary. Primary data is collected from first-hand experiences, including interviews, reviews and etc. Thanks to Airbnb, they collected this primary data, so this study get to have the secondary data. Since Airbnb has done some of the data cleaning before they publish, the data cleaning step has greatly reduced. Since they have the primary data, they had all the control to refine and publish both `detail` and `summary` data. Now study gets the benefit by utilizing this secondary data, yet some of data cleaning and sub-sampling of the data is needed since the provided data is not specifically collected for the problems proposed.

In order to clean the data, OpenRefine (version 3.4.1), an open-source desktop application is used here. Overall, the text data or string type of data was cleaned a lot here.

A. `listings_detail.csv`

In the `description` and `neighborhood` field of the data has some html tags included. For example: "Beautiful, spacious skylit studio in the heart of Midtown, Manhattan. <br /><br />STUNNING SKYLIT STUDIO (...omitted...) TOWELS<br /><br /><b>". In order to handle this, some manual assessments were done, and replaced these with an empty string.

The field of `host_response_time` has divided into subcategories: "a few days or more", "within a day", "within a few hours", 'within an hour", and not available ones. In case of not available ones, they have both blank cells and the one marked with "N/A", so merged them into a blank cells, to have a consistency in data.

The numerical field inputted in string type is also type-casted here. `host_response_rate` and `host_acceptance_rate` had a percentage unit (%) at the end of each value, so removed it. To utilize `dropna()` to remove empty cells easily, the text-filled "N/A" cells are also emptied.

There is another text entry but can be converted into numerical field. It is `price` field, and it has a character of a dollar sign ($) in front. The dollar sign has removed. The metadata indicates that this field is "daily price in local currency", which means all the price will be in USD -Airbnb have already performed a financial unification. It is safe to remove the dollar sign here, and regard all the values are in the currency of USD. There is another issue with the `price` data. There are some fields with a price of $0, so dropped these 36 listings since this seems faulty or posted by mistake. This `price` field has another issue. Since it was originally in string type of field, it has a comma (,) in its thousands. In order to parse this string into an integer or floating point values later on, this comma has been replaced with an empty string. This field also has a decimal point to indicate values less than a dollar, but having this decimal point will not be a problem when converting it into a numeric format.

To keep the data set concise and reduce the loading and processing time, the unnecessary columns have been removed. The following columns have removed: `listing_url`, `picture_url`, `host_url`, `host_thumbnail_url`, `host_picture_url`, and `license`. Additionally, `host_neighborhood` and `neighborhood` field got removed too because this field suspected to have lots of data with different naming conventions. Note that Airbnb have already provided a cleaned and unified version of this neighborhood data in `neighbourhood_cleansed` and `neighbourhood_group_cleansed`. These fields have no duplicate values, or different format with a same meaning.

There is a suspicious field as well. In `bathrooms_text` field, there are 2 samples which indicate that they have 10 and 10.5 baths. This seems to be an outlier and unrealistic, so manual check was done. It turned out that hosts rent the entire townhouse and apartment which sounds about to be right. They also indicated a similar number of bedrooms.

B. `reviews_detail.csv`

Similar to the `listings_detail.csv`, this review data also contains some html tags in the `comments` string field. For example: "Hi to everyone!<br/>Would say our greatest compliments to Jennifer, the host of Midtown Castle." In order to handle this, some manual assessments were done, and replaced these html tags into an empty string.

The reviews contain some foreign languages as well. In this study, only the English reviews will be utilized to yield a result. Since language data cannot be filtered by OpenRefine, it will later be filtered and dropped in Python script.

According to their metadata, all the date-related fields are in a type of `datetime`, which is a standard object to interpret a time data. There are multiple strategies to handle missing fields in data. This includes simple removal, substitution, forward and backward fill for the time-series or continuous data, and impute. In this study, note that whenever examining the data, the missing fields have removed with `dropna()` first before it gets analyzed further. Even after the removal, the sample size is still greater than 20,000, which is the bare minimum here to represent the whole population.

IV. Exploratory Data Analysis (EDA)

Exploratory data analysis, often shortened as EDA, tries to maximize the insight into a data set and into the underlying structure. EDA may give insights of outliers, factors that affect significantly or minute way, ranked list of factors by their importance, and etc. Throughout EDA, one can examine whether to go back to the data cleaning step and perform more cleaning or not. In this section, both graphical and non-graphical EDA will be performed for an analysis.

A. Non-Graphical EDA

Since this study is focused on the users' reviews and the price of the listings, the non-graphical EDA has performed on those two criteria.

As seen in Table I, it shows a tendency that people are used to be satisfied with their place. The mean value is over 4.5, and even the 25% value is also over 4.5. Yet, we should keep

TABLE I
REVIEW_SCORES_RATING (0-5 SCALE)

```
count    27627.000000
mean         4.578315
std          0.854467
min          0.000000
25%          4.570000
50%          4.820000
75%          5.000000
max          5.000000
Name: review_scores_rating, dtype: float64
```

in mind that the minimum was 0.0, and there are some people who got really frustrated with the place. Another thing to note is that the number of total sample size is 27,627, which is an adequate amount of samples to perform an analysis. Note that the empty cells have dropped beforehand.

TABLE II
PRICE (USD)

```
count    36887.000000
mean       169.351126
std        299.216651
min         10.000000
25%         70.000000
50%        110.000000
75%        185.000000
max      10000.000000
Name: price, dtype: float64
```

The second non-graphical EDA was done on `price` of New York City's Airbnb listings. The result in Table II indicates that the listing has a mean price around $169.35, and it can be as expensive as $10,000. So far this price seems like an outlier, but it is difficult to determine that without examining the graphical plot with where they are located. Thus, this needs further study. Note that the prices with $0 and empty cells have dropped before the analysis. The minimum price, a dollar, also seems valid since those listing have reviews on them. In addition, the number of total sample size is 36,887, which is enough to perform an analysis on.

B. Graphical EDA

Due to the limitations of non-graphical EDA, graphical EDA have also been performed more extensively. Non-graphical EDA is quite limited since it is less intuitive and difficult to visualize the distribution and outliers of the data.

The first graphical EDA was performed on reviews, in a box plot format. As seen in Fig. 1, the distribution is chunked in the review rate of 4.5, and Q1 and Q3 values are also located there. The red-colored are indicated as its outliers, but it cannot be definitely concluded that they are outliers; people might be really frustrated or disappointed about their places. However, the review with 0 looks like an outlier or erroneous data. Thus, manual validation was performed, and it is concluded that these 0-rated reviews seem to be valid, based on other sub-review categories like cleanliness, communication, and etc.

The price box plot is more interesting. The Fig. 2 shows that the majority of prices are distributed in a range of less than $500, but it has a lot of outliers at the same time. There are also two distinct outliers with a very expensive price per night. These have further examined, and concluded that they are not outliers. Some people actually have stayed there and wrote decent reviews. Still, there are a lot of outliers, so this would be examined through histogram distribution.

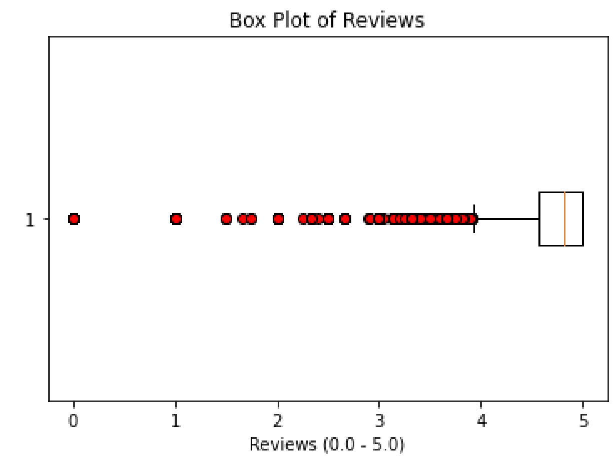

Fig. 1. Box Plot of Reviews In New York City's Airbnb

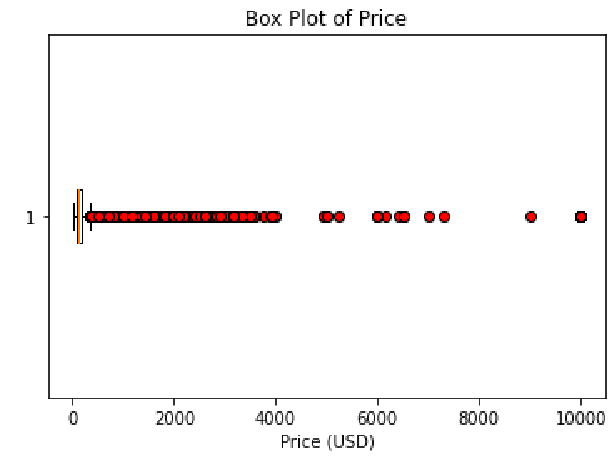

Fig. 2. Box Plot of Price In New York City's Airbnb

In Fig. 3, the further examination about the price distribution was performed. This price distribution is significantly skewed to the right. Since the majority of distribution is congested in prices less than a thousand, considering a subsampling of this data will also be meaningful and drop the prices which are greater than $1,000. Yet, the tail cannot be concluded as an outlier since they were renting their entire house, have extra features like hot tub, and etc.

To further analyze the price distribution, the data set has visualized based on its county in Fig. 4. It is clear that Manhattan has the high price listings, compared to other county. For instance, Bronx and Queens tend to have low prices since they are more of suburb and close to Long Island, not the central area of New York City.

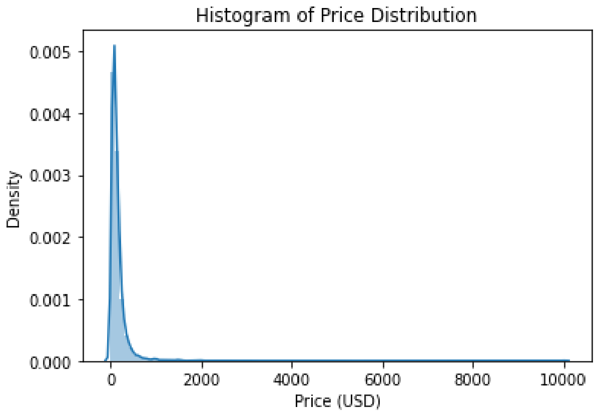

Fig. 3. Histogram of Price Distribution In New York City's Airbnb

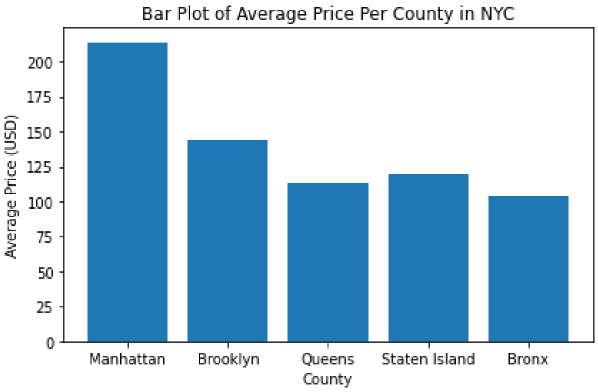

Fig. 4. Average Price Per County In New York City's Airbnb

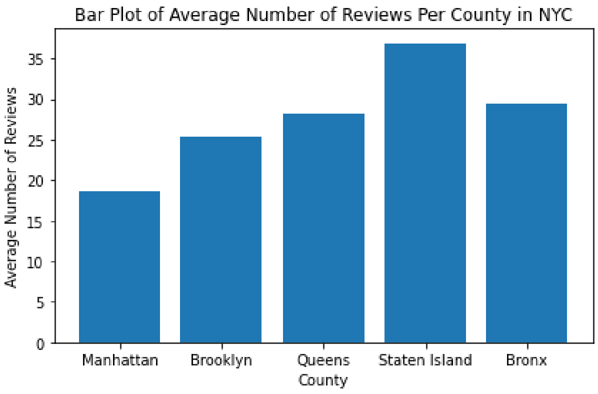

Fig. 5. Average Review Per County In New York City's Airbnb

To examine how many reviews that users have written after their stay per county, another bar plot on average number of reviews have done in Fig. 5. Interestingly, Manhattan reported the lowest number of average reviews. An indication can be made here -since the price is too high, less people might have stayed here. It would be better to cross-validate this hypothesis, but this provided secondary data did not have an information about number of users have stayed. In case of Staten Island, which reported a decent amount of average price, has the most number of reviews written. People might have stayed in this county a lot since the value is reasonable.

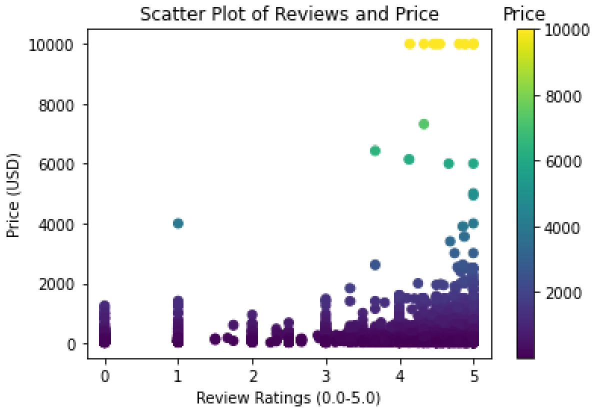

Fig. 6. Relationship Between Review & Price In New York City's Airbnb

In Fig. 6, an analysis was done to examine the relationship between reviews and prices. Interestingly, even if the prices are high, people still give high rates on their reviews. There might be various reasons behind, one can be that the place was clean, or had decent amenities. Since the majority of the listings have a price of below $1,000, reviews are congested in those range. To conclude, regardless of the price that they have, the review varies.

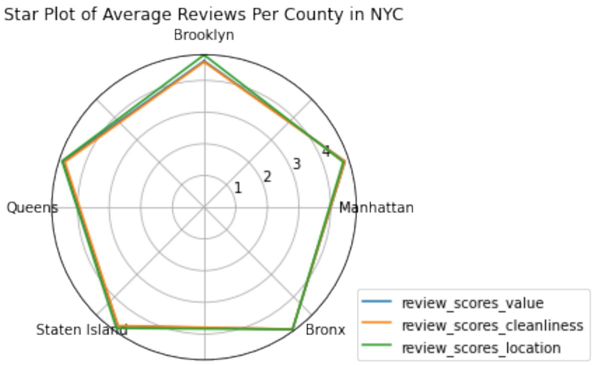

Fig. 7. Different Reviews Per County In New York City's Airbnb

In the provided data set, it has a different category of reviews. Aside from overall review score, it also has a reviews of accuracy (how accurate the description is), cleanliness, check-in experience, communication with host, location, and its value. Out of these available reviews, three of them were chosen, which are value, cleanliness, and location. Since most of the reviews reported an average rate of 4.5, it does not really depict the interesting changes here. Out of these, Brooklyn reported that although their location is decent, the cleanliness is a little poor. In general, users seem to be overall satisfied during their stay.

In this section, multi-dimensional EDA was performed to move onto the further analysis of the data. EDA is an important step in data science pipeline to have a concrete understanding of the data, including understanding outliers, variable distributions, and relationships between variables. During this step, some of additional data cleaning was concurrently performed with a manual validation.

V. Statistical Model

To have a further understandings of the data, statistical model is often used. This model allows to summarize the statistical feature of the data set. In data science pipeline, it is a taboo to utilize the whole data set; while the subset is used. It is not only difficult to get the real-time changing and updated data every time, but it may also be erroneous so data cleanup is needed all the time. If the sampled data has decently sampled, although the population is small, it may represent the total population. In this sub-sampled data, statistical model has an ability to indicate whether there is a certain significance level in the data or not. In this section, the relationship of data set has been examined through correlation and hypothesis testing.

A. Correlation

Correlation coefficient is a good measure to get an insight of whether two variabled are related or not. There are two types of coefficients that will be discussed here: Pearson correlation coefficient, and Spearman Rank correlation coefficient. Pearson correlation coefficient measures the nature and strength between two variables, but even a one single extreme value may affect the coefficient. Even the case where the length or size of the two variables are different, Pearson correlation coefficient may handle it. However, Pearson correlation coefficient is limited in case where the data is non-linear or has a lot of noise and outliers. To overcome this limitation, Spearman Rank correlation coefficient is used. In Spearman Rank correlation coefficient, it does not fit into a line; rather, counts the number of disordered pairs, and focus on their ranks. The value of coefficient ranges from -1.0 to +1.0. The sign indicates whether two variables are postively or negatively correlated, and the higher the value, the more they are correlated. Yet, these correlation methods may fail to analyze the relationship between two variables. Sometimes, examining their cluster is required. Also, note that correlation merely indicates the relationship in between variables, and cannot explain the causation in between. In this study, both correlation coefficients are computed to examine any type of potential relations in between.

TABLE III
Correlation Coefficient Between Host Response Rate And Review Scores On Communication

Pearson	Spearman Rank
0.1299935274550822	0.1372105386148191

In Table III, the correlation coefficient is computed for two variables of host response rate and the reviews on communication experience. The underlying assumption was that those two variables are related because if the host replies back in a prompt manner, the better the communication experience that the user gets. However, both Pearson and Spearman Rank coefficient values reported that there is a negligible correlation. Thus, the assumption was incorrect, and the response rate and the review scores on communication is barely related.

TABLE IV
Correlation Coefficient Between Host Acceptance Rate And Number of Reviews

Pearson	Spearman Rank
0.1281814686462352	0.16401471577448987

Another correlation coefficients were computed on host acceptance rate and the total number of reviews in Table IV. The assumption was that the more the host accepts the people to stay over, the more the review that they gets. However, it reported that they have a negligible positively correlated trend between two variables. One interesting finding to note is that Spearman Rank correlation coefficient reported a slightly higher value, so these two variables are rather more of non-linearly correlated.

B. Chi-Squared Test

In χ^2 test, it sets up a hypothesis and check for the significance in its evidence. In hypothesis set up stage, the research question or assumption has converted into a null hypothesis and the alternative hypothesis is when the null hypothesis is false. Afterwards, apply the test statistic and get the p-value. Lastly, compare this p-value to the significance level and draw a conclusion. χ^2 test is used to test the independence of two variables. In this study, multiple χ^2 test has proposed to analyze the independence between two variables.

The first test was performed on the superhost and the price of the listings. The null hypothesis here is that the superhost and the price of listings are independent, and the alternative hypothesis is that it is difficult to conclude that whether two variables are independent without more data (two variables are dependent). Firstly, the model differentiated the binary value of superhost into whether it is true or false. Then for the price the threshold has set to the $1,000, and regarded the listing with $1,000 and over is the expensive one. The threshold value has driven from Fig. 3, which depicted that most listings are under $1,000. The p-value resulted $\approx$ 0.0099, and the significance value was set to 0.05, the most widely used one. The null hypothesis gets rejected here since the p-value is less than the significance level. Thus, superhost and price of the listings are independent, and the price varies regardless of whether the host is superhost or not.

The second χ^2 test was performed on the superhost and the rate of reviews. The underlying assumption was that if one is a superhost, he gets higher review scores. The null hypothesis is that the superhost and the review scores are independent, and the alternative hypothesis is that it is difficult to conclude that whether two variables are independent without more data (two variables are dependent). Similar to the former χ^2 test, the model differentiated the binary value of superhost, and then set a threshold on the review scores. In this study, the review is regarded as "high" score only if when it is higher than 4.8 -this threshold was taken since the mean value of the review score was high in Fig. 1. The reported p-value is $\approx$ 1.80e-267, which is a minute value compared to the significance level. Thus, the null hypothesis gets rejected, where the superhost and review rates are independent, and users give high reviews regardless of the fact that the host is a superhost or not.

The last χ^2 test has performed on the price and the review scores. The assumption was that the cheaper the place, the bad the review. Since there are some good reason behind for a product being cheap. The null hypothesis is that the price and the review scores are independent, and teh alternative hypothesis is that it is difficult to conclude that whether two variables are independent without more data (two variables are dependent). The threshold for price is set to \$110, based on the 25th-percentile in Table II, and the review score less than half (=2.5) is considered as a bad score. The resulted p-value was $\approx$ 4.80e-06, which is less than the significance value of 0.05. Thus, the null hypothesis gets rejected, where the price and review variables are independent.

C. t-Test

The workflow of student's t-test is similar to the one of χ^2 test. The hypothesis has been set up, and perform calculation to see whether to reject and fail to reject the proposed hypothesis. In t-test, both one sample and two sample can be used to be tested. It compares the mean value of the variable, with a hypothesis of whether the variance is same or not.

In this study, the t-test has applied to the number of listings of Bronx and Manhattan. The insight behind is that since Manhattan is a relatively expensive region compared to Bronx, hosts will more of having a singular or few listings on Airbnb, compared to the hosts from Bronx. The null hypothesis is that the average number of listings in Bronx is greater than the one of Manhattan, and the alternative hypothesis is that it is difficult to conclude that whether Bronx has a greater average listing values without more data. The p-value turned out to be 1.0, and it is greater than the significant value, which is 0.05. Thus, the null hypothesis failed to get rejected and it is difficult to conclude that whether Bronx has a greater average listing values without more data.

VI. Machine Learning Model

Machine learning studies the input and output of the data set, and gives a model to produce a high-accuracy output once the input is provided. This technique is widely used when there is a limited expertise of human being. Through this learning algorithm, multiple tasks including pattern recognition, anomalies detection, and prediction can be done. There are multiple learning mechanisms, which are supervised, unsupervised, and reinforcement learning. Supervised learning provides a desired output, while unsupervised learning does not have an information of outputs. In case of the reinforcement learning, the model gets reward after taking each action. This study will mostly cover the supervised learning, by taking the benefit of provided data set.

In this study, three machine learning models are proposed to perform a binary classification to determine whether the host is a superhost or not. The three types of models are: k-nearest neighbors, logistic regression, and gradient boosting classification. k-nearest neighbor calculates the distance between features and get the nearest neighbors around. It is fast since it does not require any training, but its distance metrics is important since it simply relies on the distance between samples. Logistic regression is another useful technique in binary classification. It maps the regression value from $(-\infty, \infty)$ to the range of [0, 1] using a logistic function. Logistic regression also takes a relatively short runtime to be trained. Gradient boosting classifiers stack different machine learning models together to create a strong predictive model. This iterative process causes a longer runtime, but the boosting and stacking mechanism makes the model to be more robust. When applying these models, cross validation was also used to prevent the model being overfitted. The result also provides an average of k-fold cross validation score.

TABLE V
SUPERHOST CLASSIFICATION MODEL WITH HOST FEATURES

	kNN	Logistic Regression	Gradient Boosting Classification
CV Accuracy	0.7284	0.7383	0.7459
Test Accuracy	0.7205	0.7245	0.7329

The first scenario is to determine whether one is a superhost or not, with features of host response and acceptance rate. This model relies more of host's own feature. Each model's scores can be found in Table V. The result indicates that there is not much difference among three models, but the gradient boosting classification model performed the best by combining multiple models. As expected, the test set accuracy is lower than the trained k-fold average accuracy score.

TABLE VI
SUPERHOST CLASSIFICATION MODEL WITH HOST AND OTHER FEATURES

	kNN	Logistic Regression	Gradient Boosting Classification
CV Accuracy	0.6964	0.6748	0.7859
Test Accuracy	0.6802	0.6683	0.7780

The second model has more features included. In addition to the host's own features, the total number of reviews, review rates, and price features have included. Having more feature may provide a more information to the model and aid its accuracy, but it might not be a right choice of doing it and may have to reduce its dimension. As seen in Table VI, kNN and logistic regression model reported poor accuracy compared to the simpler model with less dimensions. On the other hand, gradient boosting classification model reported the best accuracy value here. Thus, it is better to perform a dimension reduction for kNN and logistic regression models to predict superhost. While gradient boosting classification model may take larger dimensions of data and perform better through stacking its models.

In general, k-nearest neighbor model reported the most poor accuracy result on both cross validation and test set. The accuracy of this kNN model is expected to be increased with a larger size of data, but it is not available at this point.

Another machine learning models have proposed to predict the price of listings here. Linear regression and gradient boosting tree models have used to predict a price. Linear regression model is simply a line-fitting problem, which is looking for a best linear function to explain the data. Similar to the gradient boosting classification model, gradient boosting tree consists of many trees, and each of them gets trained on the residual of the previous tree.

TABLE VII
PRICE PREDICTION MODEL

	Linear Regression	Gradient Boosting Tree
CV R^2	0.0408	0.0408
Test R^2	0.0303	0.0406

The result in Table VII shows that both models are poorly predicting the price of listings. To examine its performance, R^2 score was used to quantify the performance of these regression models. This R^2 score reported around 0.04. This indicates that the features failed to explicitly represent the price, nor closely related. Yet, gradient boosting tree's performance revealed consistency in both training and test set, unlike the linear regression model.

VII. Conclusion

Several multi-dimensional analysis including exploratory data analysis and statistical models have been proposed on New York City's Airbnb data to explain the relations between each feature. Although multiple assumptions have made, the result indicates that those features are not quite related. The machine learning model to predict a superhost reported a decent performance, yet still needs an improvement on its accuracy. Price prediction model failed to predict the price with given features. Making the model more complex via enlarging its dimension should be further considered.

VIII. Future Work

The result clearly shows that the improvement in price prediction machine learning model is needed. Instead of applying a simple machine learning model, using a layered neural network should also be considered. In order to analyze the given data in multi-dimensional way, further research should be done in the future. The future work includes more analysis on language processing, including natural language processing, text mining to generate keywords from reviews, and sentiment analysis.