Analysis of Airbnb Pricing Level in NYC

Ju-Eun Kim - 1005935552

June 21, 2021

Abstract

Airbnb is launched in 2008 to expand on travelling possibilities and experiencing the world (Dgomonov, 2019). Many travellers around the world find it helpful for the long-time travelling budget-wise and high accessibility. However, due to the variety of the host, the location, room type, pricing, and availability are all different. The observed data were collected publicly available information from the Airbnb website (Inside Airbnb, 2021). The collected dataset describes the Airbnb listing activity and metrics in New York City for 2019 (Inside Airbnb, 2021). For this analysis, the primary purpose is to develop a guideline database for the travellers to choose a certain quality of the room at a specific budget level alongside other factors. The methods that will be used are propensity score matching and regression discontinuity. The result will provide how much budget is required for a particular type of room or the room's location in NYC and how the number of reviews is related to the price. The conclusion extracted from the result will be a valuable source for both the travellers to NYC and the hosts to manage the price level. The travellers can find a room in NYC with appropriate pricing, and the hosts can rent out the room at the best price for the travellers can satisfy.

Introduction

Airbnb was launched in 2008, and it is now a global platform for travellers to find rooms instead of expensive hotels (Airbnb, 2021). It is now almost thirteen years that Airbnb's site lists 81,000 cities globally with more than six million rooms, apartments and houses (Sherwood, 2019). Also, there are almost 4 million hosts across 220 countries (Airbnb, 2021). Every night, more than two million people choose to stay at Airbnb property these days (Sherwood, 2019). The best advantage for Airbnb is that it is cheaper than the hotel that can help people reduce their budget. Also, it is helpful for the long-time travellers that they can potentially stay for a more extended period than other facilities (Airbnb, 2021). Furthermore, unlike a hotel, most of the rooms include a kitchen. Cooking meals in the kitchen can save the budget for the travellers as well.

The main focus of the analysis is Airbnb in New York City. NYC is one of the largest cities in the US, an attractive destination for sightseeing tourists to visit famous tourist places. In 2018, there were almost 65.1 million visitors, an all-time high (NYC data, 2021). It indicates that the majority of people have used lodging services, such as hotels or Airbnb. The usage of Airbnb in New York City is high for the short-term rental for more extended travelling to NYC. As many people demand the supply of Airbnb, there are a variety of Airbnb rooms that can be found. People can have chosen rooms from various options that can be found on the Airbnb official site.

However, there is one primary concern for the Airbnb platform. There are almost 4 million hosts worldwide and more than 40000 hosts in New York City (Airbnb, 2021). Due to the variety of the host, the location, room type, pricing, and availability are all different. For example, some hosts may only provide a private single room, and even with the same budget, some may offer the whole apartment. Also, if the travellers visited NYC for the first time, it may not be evident how to choose a room due to various prices. Therefore, it is critical to analyze the data for the travellers' convenience and the fair rating for the hosts to prevent overpricing.

Hypothesis

At this moment, it is unsure if the same budget will provide the same quality or type of room. The main research question throughout the analysis would be that if the same budget will result in the same satisfaction of the room. The satisfaction is assumed to be directly related to the type of room that people get. Then, the people will leave a review about the room. Thus, the review of Airbnb will be a key to choose a room. The hypothesis is that the type of room and the number of reviews are the main factors that change the room's price. Therefore, it would be efficient to look at the observations with a similar budget is highly likely for the travellers to find the satisfying quality of the room.

Goal

The analysis will provide a guideline that the travellers can follow to find an appropriate room using various factors. Using the data found from Airbnb, the best method of finding the room in NYC will be presented. It will help the travellers to recognize if the rooms are overpriced or even in a good deal. It will be an excellent guideline for first-time travellers to NYC.

Terminology

Observational Data: The data is collected by observing certain variables without controlling specific variables. The data will be used to determine if there is any correlation. (Caetano, 2021)

Sample: THe group of observations that are actually observed.(Caetano, 2021)

Data

Data Collection Process

The data, `AB_NYC_2019.csv`, is collected from publicly available information on the Airbnb official website (Inside Airbnb, 2021). All the information was publicly allowed to see that the Inside Airbnb could collect all the Airbnb information provided on the official site. The population of the dataset is the entire Airbnb information in NYC. The target information would be all the Airbnb in NYC. Then, the target population is Airbnb that is officially on the website to provide information. However, not all Airbnb may not have all the necessary information. Some may be outdated, and some may not provide complete information. The Airbnbs that full information was available was collected, which is the sampled population.

The collected Airbnb data is the sample of the available rooms in the year of 2019 information. There may exist foreseeable drawbacks, such as quitting Airbnb's service as a host. Even if the host quits the service, the data will not be recollected for the analysis. Therefore, in the future, there may exist some fluctuation which may cause misleading for the readers. Also, the host may change the price of the rental. Then, the analysis has limitations because the result extracted will be different from the current reality. The limitation of the study is that it cannot follow the time series. It will take time to update the newly obtained information, and the past analysis may become irrelevant at a fast pace. In addition, there is possible survivorship bias. The hosts that managed to maintain the room will be the only ones included in the analysis. Even though it is the analysis for the year 2019, if the host quit the business in early 2019, they may be omitted, thus causing bias by not providing enough information (Caetano, 2021). Also, some hosts did not give enough information for the analysis, which may cause self-selection bias, which is also called the non-response bias (Caetano, 2021). If the hosts do not answer the question fully, they cannot be used for the analysis. However, if the observation included essential information that can potentially cause changes in the model, it will cause a significant bias toward the analysis.

Data Summary

The data is obtained from `Kaggle`, originating from the `Inside Airbnb` website. The data includes the information about Airbnb in New York City with the Airbnb's name and the host's name. The private information was not collected, but only the information that was publically announced. It has information

about Airbnb's pricing, location, room type, minimum rental days, number of reviews, and availability in a year. The analysis will be done using the key section of the data. [8]

However, all data may not be necessary for the analysis. Using the whole dataset may cause confusion and increase the complexity. Therefore, the information that is irrelevant for the analysis will be cleaned out.

Also, the rooms that do not provide the complete information for the cleaned data are considered unnecessary because they cannot be analyzed. Thus, they will be removed from the dataset.

*Refer to the appendix for the overview of the cleaned dataset.

Key Variables

The `neighbourhood_group` is one of the critical variables that can influence the price. It is a categorical variable that determines the location of the room. The famous sightseeing places are crowded near Manhattan (Liza, 2017). The majority of landmarks are located in Manhattan, and the public transit system is convenient for travellers to move around the whole of NYC. Therefore, it is likely that it would be more expensive to rent a room in Manhattan. The further from the core of the city, it is likely that the price of the room would decrease. However, it will not be as convenient for travellers to move around using the public transit system. It will be one of the vital factors for travellers to consider when renting a room.

The `room_type` is also a critical factor for room selection. It is a categorical variable that determines the room type that the renters can borrow. The travellers will have a preference between the lower price and the better quality of the room. For example, the room with privacy secured will have a higher price, but the traveller may prefer to pay more to rest their heads. However, on the other hand, even if they use the shared room, they may prefer to save the budget. Therefore, there is a variety of selection that they can choose from the options. The `room_type` is highly likely that is related to the `price`. However, renting the whole apartment may not always be more expensive compared to renting the shared room. There might be other factors that play a role. The details will be discussed in the method section.

The `price` is a primary variable that both the long-term and short-term travellers consider. It is a numerical variable that there is a varied range of price. The cheap rooms can be found for under 50 dollars, and the expensive rooms are over 200 dollars. The short-term traveller may prefer to use a more expensive room in a better location while they do not have lots of time to travel around. However, long-term travellers would like to pay less so that they can save their budget. It would be best for both groups to obtain the best room that they can get within the budget. Therefore, the price will be analyzed thoroughly alongside the other factors. It will help them to find the best room that they can satisfy during the travel to NYC.

The `minimum_nights` and `availability_365` are the variables that can explain the price of each room. They are both numerical variables that can potentially explain the cost of the room. The `minimum_nights` represent the minimum number of days that the travellers have to rent. It is expected that the more days they have to stay, the lower the price per night. One of the reasons is that host does not have to clean the room every day while they are renting, and it has a shorter term to leave the room empty until the subsequent occupation of the room. The `availability_365` represents the available days of room for a year. This fact is observed because the Airbnb listings are not always on the website for 365 days. The host may decide not to lend the room in certain seasons. Then, the room may be only available for specific days. Then, it is likely that the short-term availability rooms are more expensive while the host will tend to maximize the profit from Airbnb in the short term. Therefore, the two key terms may be critical for travellers when they select the rooms.

The `number_of_reviews` will help to indicate if the specific Airbnb is popular or not. The travellers tend to make a reservation to the room by reading the review. It can determine how many people have visited Airbnb. Also, for most travellers, as mentioned, the majority of people staying in NYC prefer to use Airbnb who is running tight on their budget. If there are more reviews than the other rooms, it indirectly indicates that the price per night will not be high but relatively cheap. The more review there is, it is likely to have many visitors due to good quality and lower price than the other rooms. Therefore, it might be significant to consider when conducting the threshold between the excellent quality and cheap Airbnb rooms versus the expensive rooms.

Numerical Summary

First, the numerical summary of the price of each Airbnb room will be analyzed. [8]

mean	median	sd	min	max
152.7207	106	240.1542	0	10000

In NYC, the hotel features rooms start at 200 dollars per night in 2019 (Cross, 2019). It is assumed that the price of the average hotel is higher than the Airbnb rental. The mean price of Airbnb per night is 152.7206872 dollars, and the median was 106 dollars. Compared to the hotels, they feature a much lower price that is affordable for travellers. However, since there are a variety of hosts, the standard deviation of the price is very high. It indicates that the price level differs by a significant amount. The minimum cost for Airbnb is 0 dollars, which is provided free. In contrast, the most expensive room is featured at approximately 10^4 dollars per night. Even though there are outliers, most of the rooms are at a reasonable level that many travellers can easily approach with a lower budget according to the median of the Airbnb room.

Next, the number of reviews per Airbnb is analyzed.

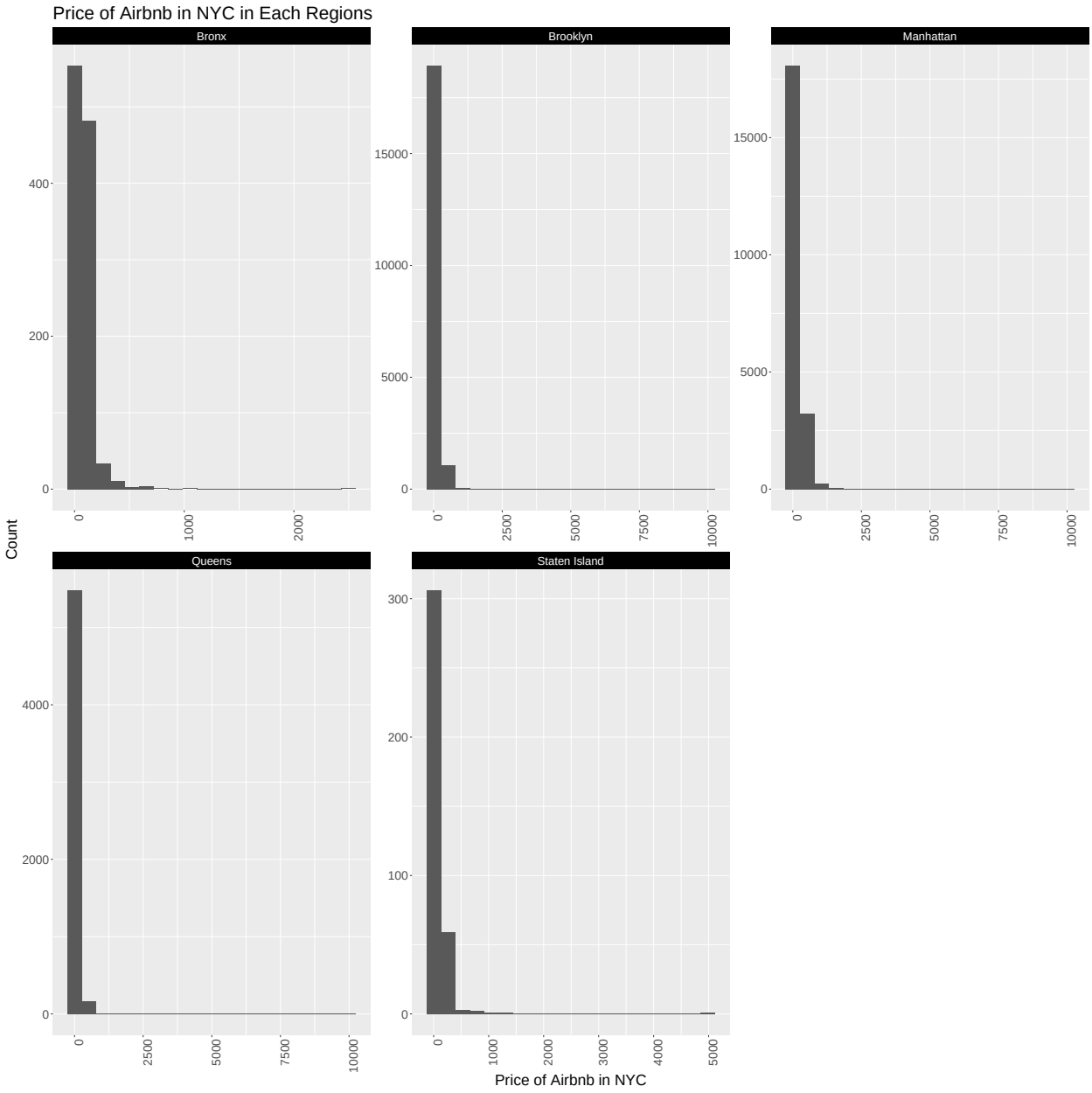
mean	median	sd	min	max
23.27447	5	44.55058	0	629

The mean number of reviews is 23.2744657 with the standard deviation of 44.5505823. It has a large spread. The median number of reviews is only 5. There are significant differences between mean and median. It indicates that the result is highly right-skewed while the mean is significantly greater than the median. Also, the maximum number of reviews that one Airbnb had was 629. Compared to the minimum number of reviews, it had a large gap. Thus, the summary table indicates that a variety of review exists in NYC Airbnb. The room with lots of reviews are more likely to be preferred, but the rooms with few reviews are not preferred while it is hard to know how the service of the particular room would be. Therefore, the result can be helpful for further analysis.

However, one concern is that the newly launched rooms may not have enough review compared to the old rooms. This fact may cause bias in the result.

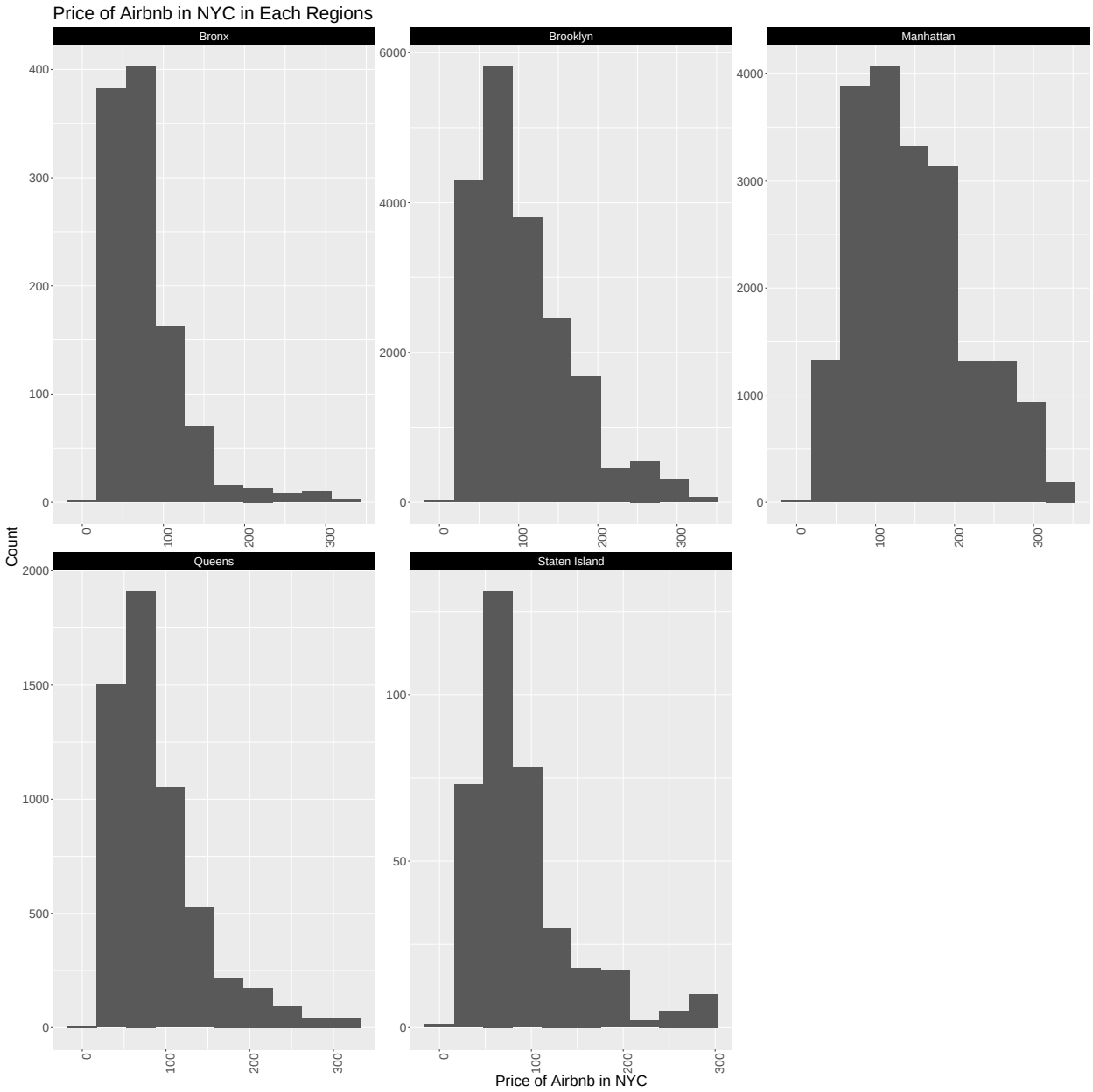
Visualization

The first plot will indicate the price of the Airbnb room. The histogram will be generated. However, it will be separated by the neighbourhood. It will indicate the price level of room in each region in the set of histograms. [8][13]



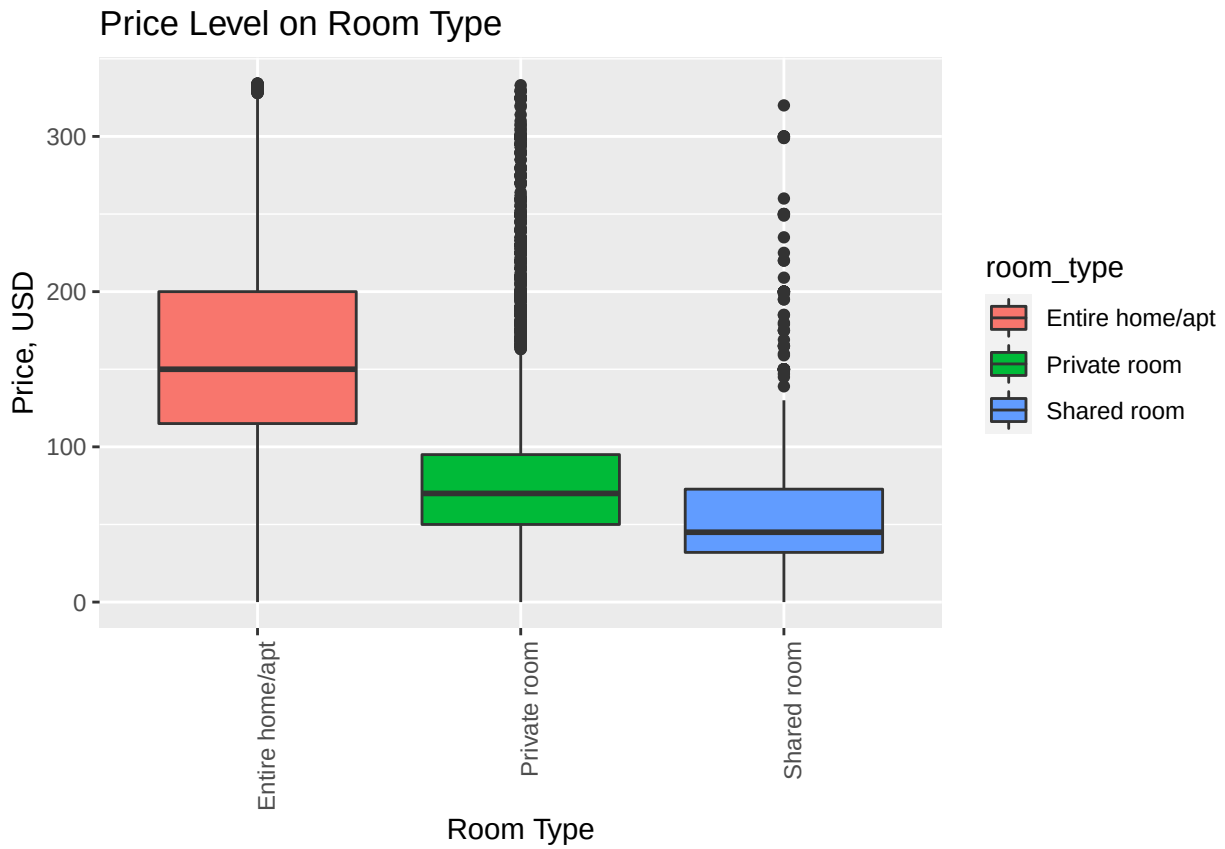
The graph demonstrated the price of Airbnb rooms in each region. The graphs clearly indicate that each neighbourhood has a similar shape of graphs. All the neighbourhoods had the most rooms of Airbnb that were under the 200 dollars. Also, there were very few rooms that were over 2000 dollars. They considered the outliers. While the median price of the room was 106, it is likely that the travellers can get the room within 150 in all the locations of NYC that they desire to stay. The other fact that can be observed is the number of rooms in each neighbourhood. By looking at the counts, the rooms in Brooklyn and Manhattan were the majority of the rooms. While the two locations are the most popular tourist place, more visitors exist and demand the room. More Airbnb rooms were developed to supply the room at a satisfying price level.

However, the graph generated may seem hard to interpret. Therefore, the outliers were removed. Then, the same set of histograms are generated again.[8][13]



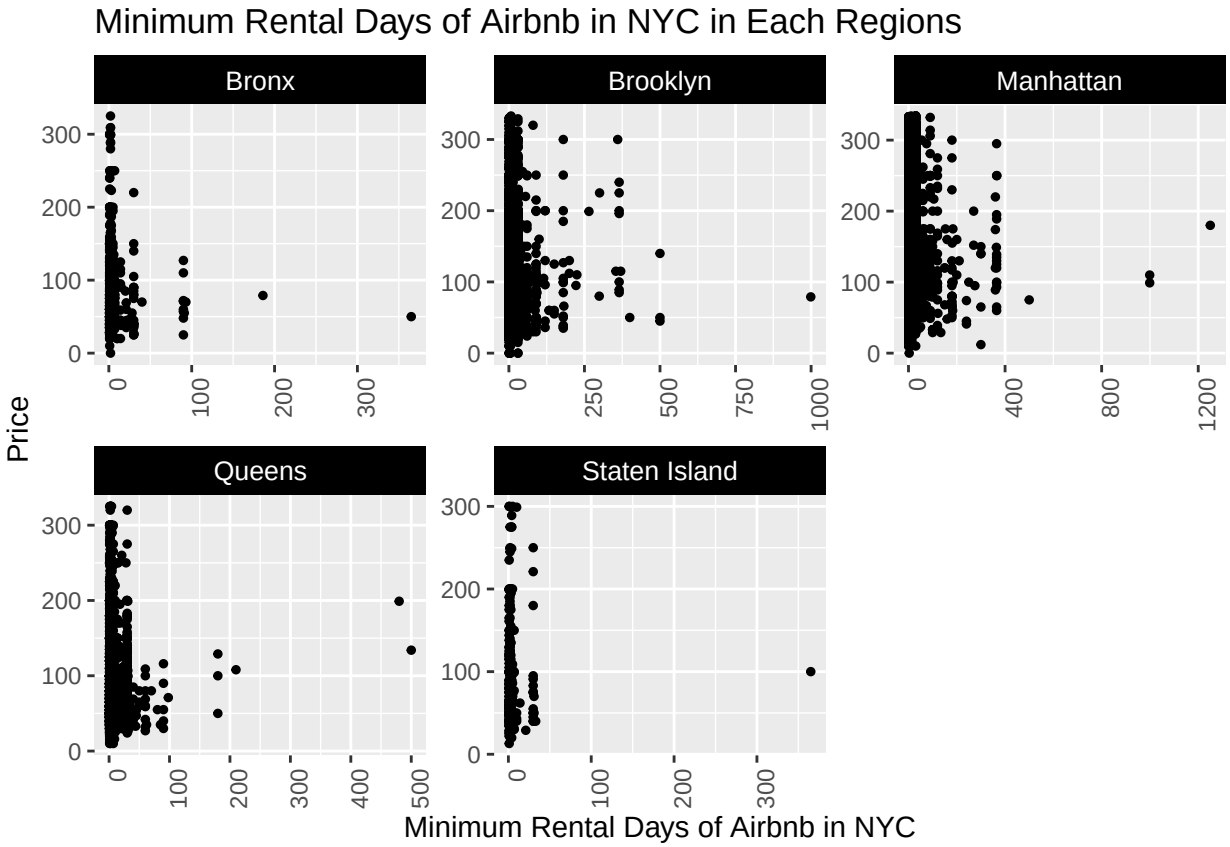
The set of graphs became easier to interpret. The mean should be greater than the median because the graphs are right-skewed. The right-skew indicates that the tail of the graph is on the right side. The mean price of the Airbnb rental is higher than the median price. Therefore, the travellers can find rooms that are cheaper than the mean easily. Also, according to the graphs, most of the rooms are around the level of 100 dollars. All the neighbourhoods provide less than 100 dollars for the travellers that are short on the budget as well.

The following graph focuses on the type of room depending on the price. The price level of each room type will be generated through the sets of boxplots for the comparison between each room type. Again, to avoid the outliers, the dataset without the outliers will be used to plot the data. [8][13] * The graph without the outlier removal is included in the appendix section.



The set of boxplots show apparent differences in the price level for each type of room. Renting the entire home/apt option was the most expensive choice. Also, the distribution of the price range was the highest as well. It ranges from 0 dollars to approximately 320 dollars, not including the outliers. The median price of the entire home/apt was approximately 150 dollars. The private room was the second expensive that the travellers could borrow. The median is less than 100 dollars, and the price range was 0 to 160 dollars, not including the outliers. The choice of range was smaller than the entire home option. Many outliers had a variety of ranges. Lastly, the shared room had cost the least amount of money. The median price was 50 dollars that which is much cheaper than the other options. There was an approximately 100 dollars difference between the median of borrowing the entire home. However, the range of choice was much smaller than the other options. Similar to the private room, there were many outliers for the shared room as well. It would be most convenient for the travellers to decide the room type depending on their budget.

The following graph is the scatter plot. It demonstrates the minimum rental days to rent the room in each region in the set of scatter plots. Again, the outliers are removed. [8][13] * The graphs without the outlier removal are included in the appendix section.



Most of the rooms can be rented in all the regions even if the travellers stay only a day. However, some Airbnb requires to stay at least 100 days to rent the room. Even though there are restrictions on renting a room, they are mostly cheaper than the other room. In the Bronx, the room requiring 365 days to stay has less rental price than the other rooms. It is similar in the other area as well. Therefore, it would be the best option for the long-period travellers to stay at a house with a higher minimum night required for the rent. The best area that they can choose would be the Bronx, which requires them the least amount of cost for a long-period contract. In contrast, the short period renting people can avoid these options instead of paying fees for extra days even when they do not stay home. For them, Manhattan and Brooklyn should be the best options. They provide the wide variety of option for the short term renters that has minimum one day contract. Therefore, the choice of regions can play an important role depending on where they would want to stay.

Methods

The two methods are selected to use as the primary methods. The first method is propensity score matching, and the second is regression discontinuity. For the brief explanation, propensity score matching is the statistical method that attempts to estimate the effect of a treatment by various factors into an account that predicts whether an individual can receive the treatment (Caetano, 2021). Next, the regression discontinuity shows whether the candidates are selected for treatment based on whether their value for a numeric rating exceeds a known, precise and non-manipulated cut-off (Caetano, 2021).

Propensity Score Matching

Overview: The propensity score matching will match people looking for the entire home/apt with a similar budget and desired location. The Airbnb that were manageable to occupy the whole home/apt will be matched, and the function that can explain whether the Airbnb was treated (that got the entire home/apt) will be obtained. Then, the forecast will be made to create matches. Every Airbnb treated, and the untreated

place was considered similar to them based on the propensity score, will be matched. The matched dataset will be left, and it will be examined to see the effect of being treated on average when they have a similar budget.

The objective of propensity score matching is to match based on observable variables. In this analysis, neighbourhood, a minimum number of nights to stay, and availability will be the observable variables.

All predictor variables will be treated as individual variables. Before creating this model, the variable named `obtain_entire` will be added to the data. This variable will have a value of 1 if Airbnb provides an entire home/apt and a value of 0 otherwise. The variable `obtain_entire` allows the creation of a logistic model.

The model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{neighbourhoodgroup} + \beta_2 x_{minimumnights} + \beta_3 x_{availablity365}$$

Then, the process of propensity score matching will assign some probability to each of the observations in the data, `AB_2019`. The probability is calculated based on the observation's values for the predictor variables at their values before the treatment (Alexander, 2021). That probability is the best guess of the observation being treated (renting the whole apt), regardless of whether it was treated or not (Alexander, 2021).

For a simple example, if the selected Airbnb chosen by a traveller with 150 dollars budget were treated but the selected Airbnb chosen by a person with 149 dollars budget were not, then as there is not much difference between the budget, the probability of being treated would likely be reasonably similar. Then, the comparison of the obtained outcomes can be made with similar propensity scores (Alexander, 2021).

The new model is:

$$Y_{price} = \beta_0 + \beta_1 x_{neighbourhoodgroup} + \beta_2 x_{minimumnights} + \beta_3 x_{availablity365} + \beta_4 x_{obtainentire}$$

After matching, the quality of matching and the outcomes will be evaluated (Caetano, 2021). Then, the examination of the 'effect' of being treated on the budget can be conducted.

The most advantageous point can be obtained from the propensity score matching to identify appropriate treatment and control groups. In this case, Airbnb that allows occupying the whole home/apt will be the treatment group. On the other hand, Airbnb with a similar budget that did not obtain the same type of room will be the control group. Also, the additional advantage of propensity score matching is that it allows us to consider many predictors at once easily, and it can be constructed using logistic regression (Alexander, 2021).

The model assumes the treatment being binary. It means that the treatment is either given or not.

Regression Discontinuity

Overview: For the regression discontinuity, the price will be examined based on the number of reviews. People who tend to use Airbnb are primarily short in budget or people who like to travel for a more extended period. It means that more people who prefer cheaper room. Therefore, the hypothesis is that the more review there is, it is likely the more affordable room with a threshold of the number of reviews. When the number of reviews passes the specific number, the room's price is likely to become lower. This is because it would be efficient for the travellers to find a suitable room when they arrange the room with more reviews to fewer reviews. The regression discontinuity will be used and will be shown visually through the graph to prove if this hypothesis is correct. The threshold will be fixed as 350 of the number of reviews.

The regression discontinuity method is a valuable method to examine the continuous variable with a threshold to the x-variable that determines if it gets the treatment (Alexander, 2021). For this analysis, the price of an Airbnb room would be a forcing variable, and the cut-off for the number of reviews from the travellers to a certain Airbnb would be a threshold. The treatment can be determined by the forcing variable (Alexander, 2021). The difference in the room price expects to be at least 10 dollars. Otherwise, they would not take the time to find a room with lots of reviews.

The model can be written as:

$$Y = \beta_0 + \beta_1 x_{reviews} + \beta_2 I(X > X_0) + \epsilon$$

The number of reviews, which is the actual variable of interest, will be examined. Then, the indicator variable with Beta2 will represent whether or not the selected Airbnb will have more than 350 reviews, which is the threshold. If X is greater than X_0 (350), the Beta2 value will be added to Y. The price Y can be determined by the two predictors.

However, as mentioned in the data section, it concerns using the whole data. The newly launched rooms may not have enough review compared to the old rooms. This fact may cause bias in the result. Therefore, the number of reviews less than 100 will not be considered. They will be removed from the data. Then, it will be a better measure that can estimate the more accurate relation between the price and the number of reviews.

There are few assumptions: 1. The cut-off is known, precise and free of manipulation (Caetano, 2021). The cut-off is fixed. 2. The forcing function should be continuous (Caetano, 2021). If there is no jump after the threshold, it is one continuous smooth curve.

Results

In this section, the results of the statistical analyses will be delivered using the R program. The results of the two different methodologies are included in the report. Also, the interpretation of the result will be discussed.

Propensity Score Matching

The calculation of propensity score matching will be shown in this section. The result will be shown thoroughly to represent how the observational factors influence the treatment. Also, in the last section, the price level will be determined.

From the method section, obtaining the entire home/apt was analyzed. If the Airbnb hosts provide the entire home/apt, it is treated and indicated as 1. However, if they do not, it was indicated as 0.

Var1	Freq
0	23486
1	25409

There are 25409 Airbnb that provides the entire home/apt and 23486 that do not. Therefore, the dataset will match 25409 Airbnb with another Airbnb group that is very similar.

Next, using the model from the method section, the logistic regression model can provide the explanation of whether the Airbnb was treated (rent whole home/apt) by the constructed functions using the observational variables, which are: neighbourhood_group, minimum_nights, and availability_365. In other words, the outcome of interest is based on these variables. Then, whether or not Airbnb provided the entire home/apt will be examined. [10]

Then, the calculated propensity score will be used to predict if they will provide the entire home/apt regardless of whether or not Airbnb actually provided the entire home/apt regarding variables: neighbourhood_group, minimum_nights, and availability_365. The prediction will be added to the present dataset, `AB_2019`.

The probability of getting the entire home/apt was fitted and added to the dataset.

Then, the constructed prediction will be used for the matching process. The Airbnb that provided the entire home/apt (treated) will be matched with the Airbnbs that did not provide the entire home/apt (untreated) but the ones which had a similar propensity score compared to the treated group. [11]

After the matching process, the dataset is reduced with the leftover observations that are matched. The number of matched Airbnb observations was 46972. It may be a drawback because over 3000 data observations are lost through the process.

Now, the Airbnbs that provide the entire home/apt ratio is the same. Half of them will provide the entire home/apt, and the other half will not. As the last part of this method, examining how getting the entire home/apt affects Airbnb's price level (how much the travellers would pay) can be done using the linear regression function. [10]

	(1)
(Intercept)	20.688 **
	(6.751)
neighbourhood_groupBrooklyn	32.671 ***
	(6.789)
neighbourhood_groupManhattan	89.753 ***
	(6.799)
neighbourhood_groupQueens	12.428
	(7.195)
neighbourhood_groupStaten Island	7.904
	(13.063)
minimum_nights	0.127
	(0.075)
availability_365	0.162 ***
	(0.008)
obtain_entire	113.513 ***
	(2.028)
N	46972
R2	0.096
logLik	-319449.901
AIC	638917.802

*** p < 0.001; ** p < 0.01; * p < 0.05.

While the coefficient's signs are all positive, all the variables are positive. In other words, as independent variables increase, the mean of the price would also tend to grow. Also, the interpretation of the coefficient value is how much the mean of the price changes by a one-unit change in the predictor variable (while fixing the other variables as a constant). As the predictor variable, `obtain_entire`, shifts by one unit in the matched dataset, the mean price will increase by 113.513 dollars. A similar interpretation can be made for all the other variables. The `neighbourhood_group` variable indicates a categorical variable. The intercept in the model output is the mean response when the location of Airbnb is in the Bronx. The coefficient for `neighbourhood_groupBrooklyn`, 32.671 dollars, will be added to the intercept to get the mean response when the room's location is in Brooklyn. Similarly, when the location of the room is in Manhattan, 89.753

dollars will be added to the intercept to get the mean response. The other locations will have a similar interpretation. Depending on the location, the different prices from the result table will be added.

To add, the assumption was met while it identified whether or not Airbnb provided, which was a binary answer. It was either yes, it was provided, or no, it was not.

As a result, several factors seem significant. The significance was examined by the p-value. The location is an essential factor to consider, while Airbnb's primary purpose is to provide room for travellers. Specifically, whether or not Airbnb is in Brooklyn or Manhattan is critical while they were the major cities for sightseeing with landmarks. Also, availability is important as well for the price level. The price can be vary depending on the availability in a year. Lastly, obtaining the entire home/apt is significant in influencing Airbnb's price in the matched dataset. Therefore, the price depends highly on the type of room that the person gets.

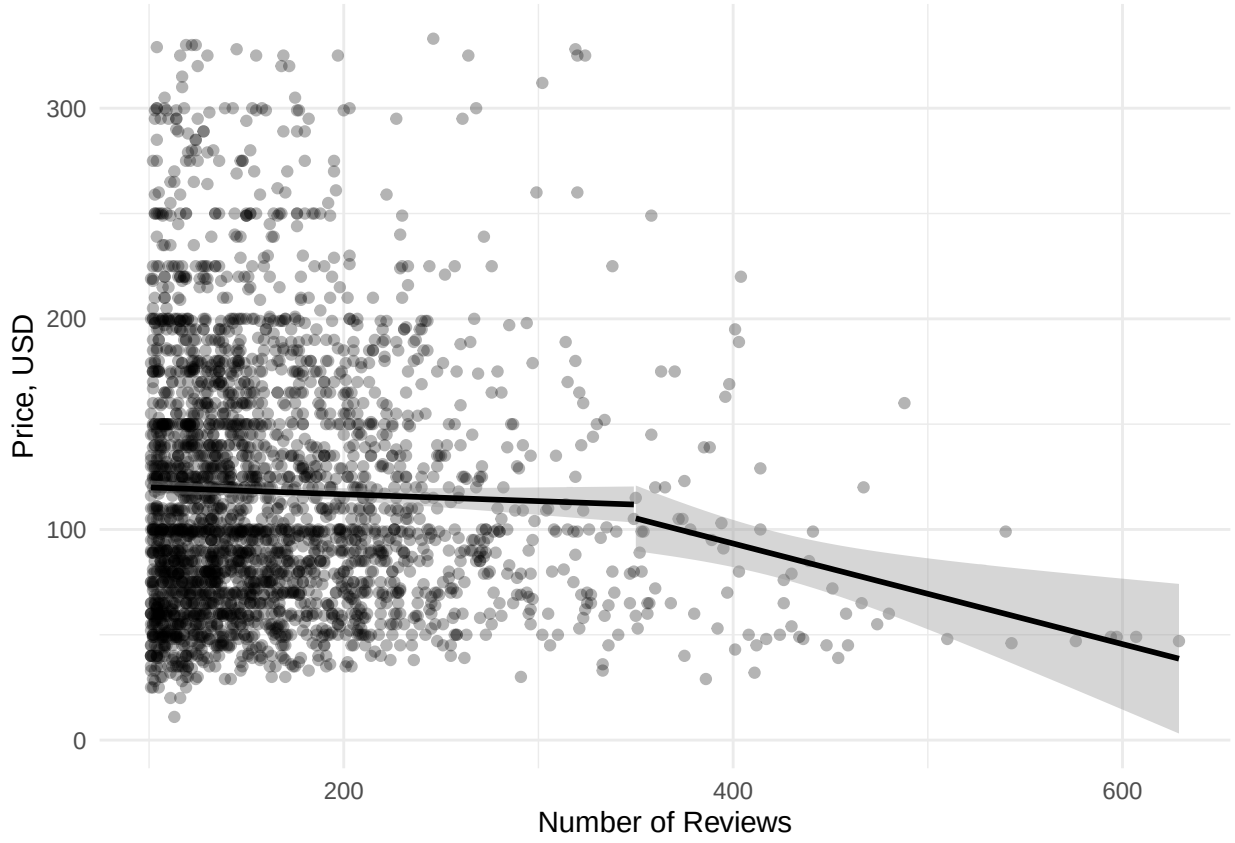
Regression Discontinuity

The visualization of regression discontinuity will be shown in this section. The difference before and after the threshold point will be examined.

Before, as mentioned in the method part, the observations with less than 100 reviews will be removed. It will result in observations that are not affected by the timeline of the launched period.

In this section, the data that does not contain the outliers will be used. The observations that are outliers of the price will be ignored for better analysis.

It leaves 628 observations in the data. Now, the data will be plotted. [8][10][13]



From the graph, it was observable that the slight jump after 350 reviews was observable. After passing the threshold, which is 350 reviews, the price of the room dropped by approximately 15 dollars. It may be a massive difference for travellers who have a tight budget. Both before and after reaching 350 reviews, it is clear that the slope between the relationship price and the number of the review was negative.

The assumptions are met while the hosts cannot manipulate the data that were already collected from the official website. Also, the threshold is fixed to 350 reviews. There are no lingering variables that could cause the same discontinuous phenomenon. Also, the forcing function is continuous while the threshold is the same for both sides of Airbnbs.

Then, the dummy variable can be used to estimate the effect of the linear regression discontinuity design. If Airbnb had more than 350 reviews, it would be recorded as `1`, and otherwise, it will be recorded as `0`. The expectation is to obtain a beta 2 value, which was based on the modelling. The increase in Y (price) intercept is expected to be around 15, as seen in the graph.

The linear regression model will be created using the number of reviews and the indicator variable (dummy variable) established in the previous step. The indicator is expected to be the amount of the jump up. [10]

term	estimate	std.error	statistic	p.value
(Intercept)	124.5627357	3.5954856	34.644204	0.0000000
number_of_reviews	-0.0410541	0.0215625	-1.903956	0.0570142
more_than_350_reviews	-18.7185856	9.1144714	-2.053721	0.0400915

The intercept of the graph is 125, which is very close to the expected value from the graph. Since the data used ignored the rooms with too few reviews, the number of reviews starts from 100 and the room with 100 reviews is expected to be around 125 dollars. The slope of the entire function is -0.0422. It makes sense while the first part of the graph was close to the slope of zero and the second part had the negative slope. While the P-value is not as small (not <0.05), the result is not statistically significant. Lastly, by looking at the indicator variable, there is an 18.7 dollar decrease in whether or not Airbnb had more than 350 reviews. As a result, they can save 18.7 dollars by choosing Airbnb with more than 350 reviews. Therefore, the travellers would look for rooms with more than 350 reviews on the official website. Also, P-value is less than 0.05, that the result is statistically significant.

Conclusions

Throughout the analysis, the main focus was on the price of the room, depending on the other factors and variables. The hypothesis was that the type of room and the number of reviews are the main factors that influence the price of the room. Two methods were used to examine the hypothesis.

The first method used was the propensity score matching method. The propensity score depending on the neighbourhood group, the minimum number of nights to stay, and the availability among 365 days was calculated. Then the logistic model where the treatment (obtaining the entire home/apt) was the outcome based on the vector of covariates was constructed (Caetano, 2021). The observations that were treated were matched with the observations that were not treated but obtained similar propensity as the treated group. Then, after balancing between the treatment and comparison group on observable traits, the outcome was evaluated. As a result, the room price per night depended on the type of room that obtaining the entire home/apt was significant in influencing Airbnb's price in the matched dataset. From this method, it was observable that the result strongly supported the hypothesis.

Next, the second method was the regression discontinuity design. The objective of the strategy was to show that the number of reviews on the official website of Airbnb can determine the price of the room. As a result, it showed that after passing the threshold of 350 reviews, there was a slight fall after for the cost of the room. Therefore, it proves that the room with more reviews is likely to have a lower price than the other rooms without reviews. As a numerical result, the room that had more than 350 reviews was cheaper by 18.7 dollars. Thus, it is beneficial for travellers to look for rooms with more than 350 reviews on the official website of Airbnb. In addition, P-value was less than 0.05, that the result is statistically significant.

The result of the analysis will be a helpful resource for both the travellers and the hosts. The travellers can decide whether they will save 18.7 dollars by making an effort to find a room with more reviews. Also, the host can recognize that the lower price is likely to result in more travellers that lead to more customers. The travellers can also plan on their budget ahead by deciding whether or not they will obtain the entire room

and consider the other factors, such as location. The result of the analysis will give Airbnb ideas to develop its business more effectively.

Weaknesses

There were few weaknesses throughout both of the methods due to the assumptions and the limitations.

For the first method, propensity score matching, there was one observable limitation. The omitted variables which were cleaned out or not used for the model may still influence the model. Then it may potentially lead to a biased result. Including one predictor variable or not including a predictor can make a lot of difference in the propensity score evaluation and matching process.

For the second method, regression discontinuity design (RDD), the jump or fall in the threshold could be modelled by another covariate (Caetano, 2021). Then, the use of regression discontinuity design is not practical. Also, the natural jumpiness should be examined thoroughly. If the depending variable Y is naturally jumpy, RDD is not the correct method to use. Lastly, the individuals should not be able to manipulate X for the threshold. If they can make a change in X, the entire design becomes invalid. However, what if the information was not correct and the X cannot manipulate? Then, it would also cause an error in the model. Thus, it can be a weakness of the method.

Next Steps

For future work, more variables must be used for more accurate analysis. In this analysis, for simplicity, only six variables (excluding the id variable) were used. However, more observation is required. For example, in method 1, the propensity score estimation was only done with the neighbourhood group, the minimum number of nights to stay, and the availability among 365 days. However, this information was not enough. More categories, such as service rating out of 5, walk score, or transit score, should be included to estimate the propensity score for matching better. Also, the matching method can be changed. The nearest neighbour matching is the simplest method that was done in this analysis. However, in the future, other matching skills, such as calliper matching and radius matching, can also be done (Caetano, 2021). Then, if they have a better quality of matching, they can be used instead on the nearest neighbourhood matched results.

For the second method, it would be better to re-collect the data by considering the time of the launched date of the Airbnb room. It was challenging to compare the number of reviews because there is a high chance that they may depend on the established date. In this analysis, the process was significantly simplified, that Airbnb with less than 100 reviews was removed from the observation. However, in this way, the data could be bias because of lots of drawbacks of the observations. Also, the old rooms likely have more reviews. In the future, if they can match and group the Airbnbs by the launching date, it can directly compare the result regardless of the timeline.

Discussion

To conclude, the price of Airbnb is one of the most crucial factors that travellers to NYC consider when choosing a room. However, other factors also take into play. Airbnb service should take the traveller's need and select the level of price thoroughly to attract more travellers to various Airbnb from other facilities, such as hotels. Unification of price level with the factors will prevent the travellers from being flocked into one specific Airbnb room. If they same similar type of room, in the same neighbourhood, and other similar conditions, it will be able to satisfy both the travellers and the hosts. The travellers can be spread out, and the hosts can have more opportunities to attract the travellers. Also, the official website should recommend the travellers leave reviews. It will help the next travellers to decide on their room and give more accurate results throughout the analysis that has been done in this paper. While the data was collected from the official website of Airbnb, providing more detailed and precise results will make the analysis more accurate and valuable for both the hosts and the travellers in NYC.

Bibliography

1. Airbnb. (2021) *The Airbnb Story*. https://news.airbnb.com/about-us/. (Last Accessed: June 5, 2021)
2. Dgomonov. (2019) *New York City Airbnb Open Data*. Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data. (Last Accessed: June 5, 2021)
3. Inside Airbnb. (2021) *Get the Data*. http://insideairbnb.com/get-the-data.html. (Last Accessed: June 5, 2021)
4. Sherwood, H. (2019, May 5) *How Airbnb took over the world*. The Guardian. https://www.theguardian.com/technology/2019/may/05/airbnb-homelessness-renting-housing-accommodation-social-policy-cities-travel-leisure. (Last Accessed: June 7, 2021)
5. NYC data. (2021) *New York City (NYC) Tourism*. https://www.baruch.cuny.edu/nycdata/tourism/index.html. (Last Accessed: June 7, 2021)
6. Liza. (2017, June 23) *GREAT SPOTS TO VISIT IN NYC*. Tripsget. https://tripsget.com/manhattan-vs-brooklyn-great-spots-to-visit-nyc/. (Last Accessed: June 11, 2021)
7. Cross, H. (2019, April 30) *New York City Hotel Rates*. tripsavvy. https://www.tripsavvy.com/new-york-city-hotel-rates-1613103. (Last Accessed: June 11, 2021)
8. Wickham et al., (2019) *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686. (Last Accessed: June 14, 2021)
9. Alexander, R. (2021, May 11) *Chapter 15 Causality from observational data*. Telling Stories With Data. https://www.tellingstorieswithdata.com/causality-from-observational-data.html#matching-and-difference-in-differences. (Last Accessed: June 14, 2021)
10. David Robinson, Alex Hayes and Simon Couch (2021) *broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/, https://github.com/tidymodels/broom. (Last Accessed: June 14, 2021)
11. Andrew Gelman and Yu-Sung Su (2020) *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.11-2. https://CRAN.R-project.org/package=arm. (Last Accessed: June 14, 2021)
12. David Hugh-Jones (2021) *huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats*. R package version 5.4.0. https://hughjonesd.github.io/huxtable/. (Last Accessed: June 14, 2021)
13. H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. (Last Accessed: June 14, 2021)
14. Caetano, S. (2021) *STA304-Bayesian-LM_Logit*. lecture in STA304, Sampling and Observational Data, University of Toronto. (Last Accessed: June 14, 2021)
15. Caetano, S. (2021) *STA304-RegressionDiscontinuity-1*. lecture in STA304, Sampling and Observational Data, University of Toronto. (Last Accessed: June 14, 2021)
16. Caetano, S. (2021) *STA304-Matching*. lecture in STA304, Sampling and Observational Data, University of Toronto. (Last Accessed: June 14, 2021)
17. Caetano, S. (2021) *STA304+-+Logistic+Regression+Intro*. lecture in STA304, Sampling and Observational Data, University of Toronto. (Last Accessed: June 14, 2021)

- The package `tidyverse` was used to manipulate and make summaries. The functions select(), mutate(), and summarize() were used.
- The package `broom` was used for regression analysis. The functions lm() and glm() were used.
- The package `arm` was used for the matching function. The function matching() was used.
- The package `ggplot2` was used to generate all the graphs, such as boxplots, histograms, and scatterplots.

All analysis for this report was programmed using `R version 1.2.5042`.

Appendix

Section 1: Supplementary Data

The table below is the original set of data, not cleaned.

id	name	host_id	host_name	neighbourhood_group	neighbourh
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensingtor
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown
3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem
3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hi
5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harle
5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hi

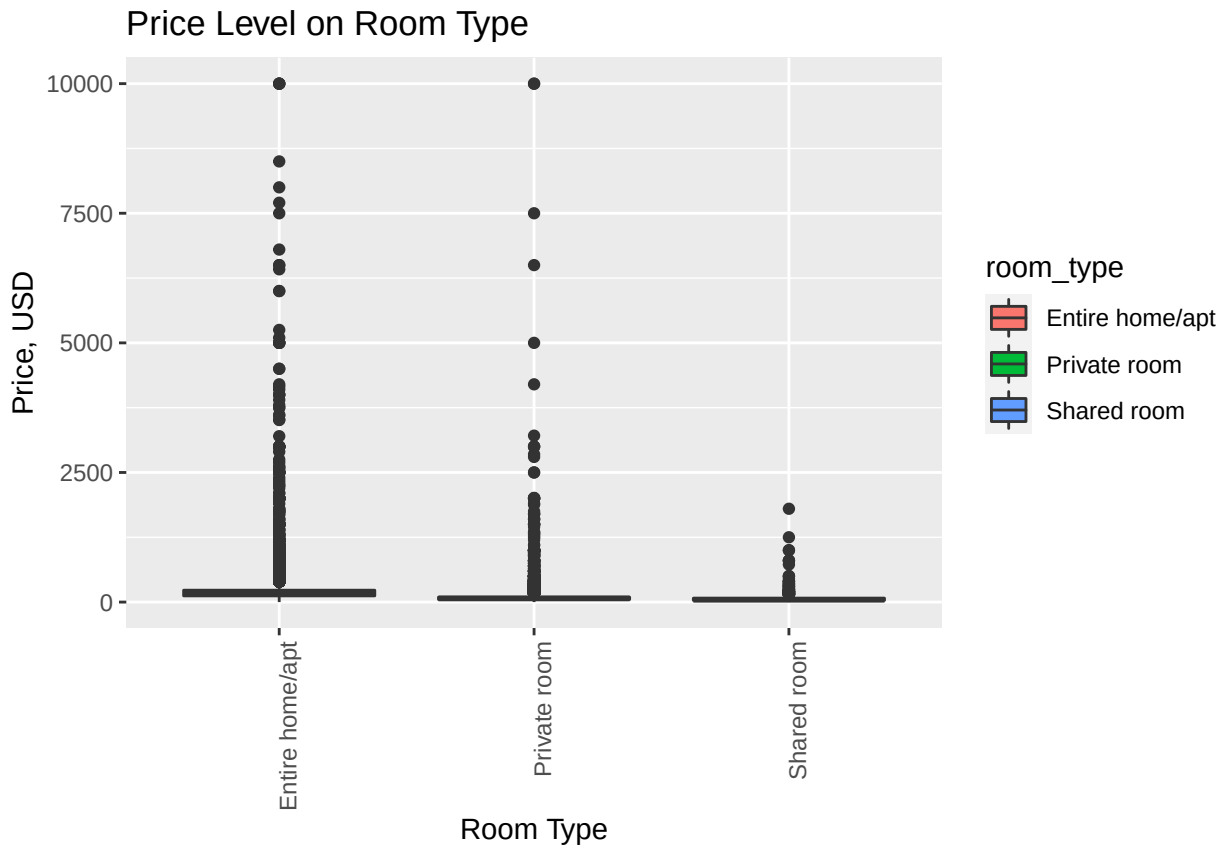
In the original data set, there were other variables such as `name`, `host_id`, `host_name`, `latitude`, `longitude` etc. However, while they were not used, it was cleaned. For further analysis, the other variables may be relevant.

The data below is the cleaned data, with all the variables used throughout the entire analysis.

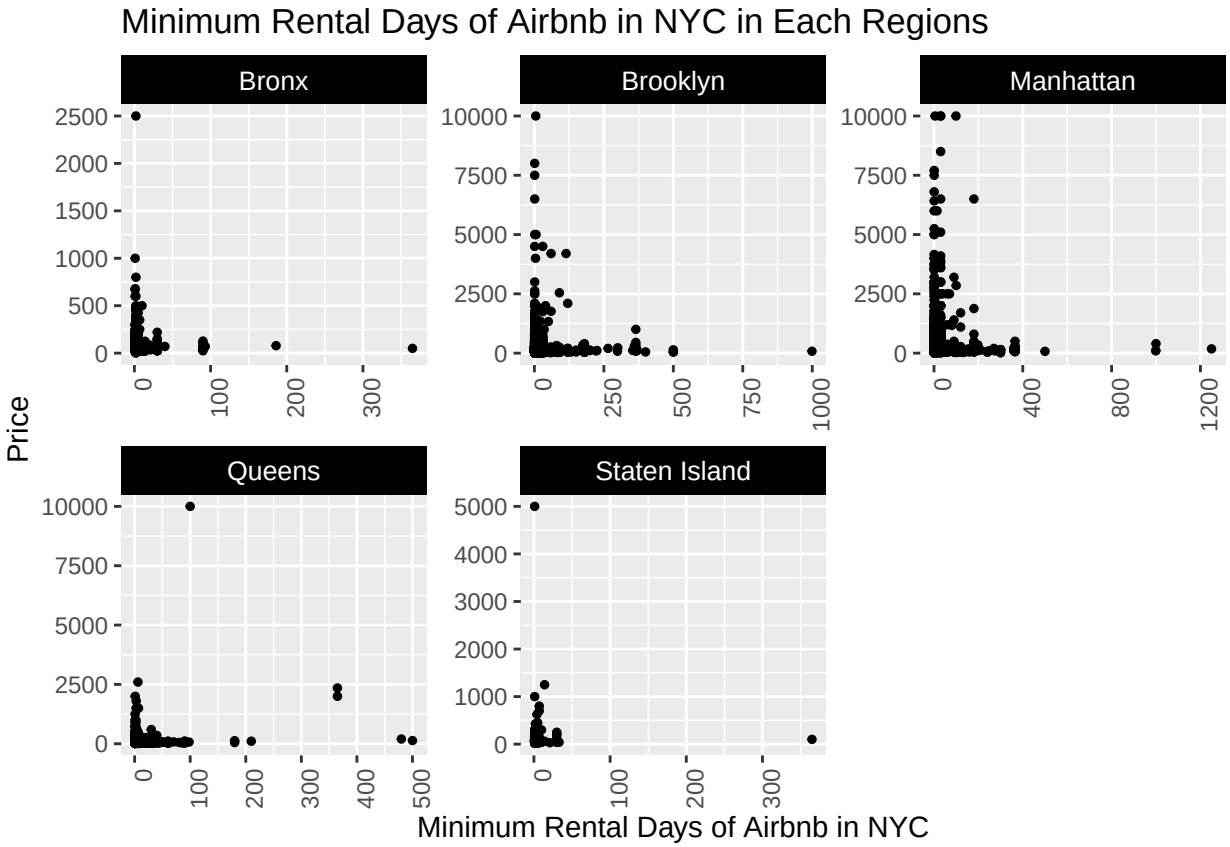
id	neighbourhood_group	room_type	price	minimum_nights	availability_365	number_of_reviews
2539	Brooklyn	Private room	149	1	365	9
2595	Manhattan	Entire home/apt	225	1	355	45
3647	Manhattan	Private room	150	3	365	0
3831	Brooklyn	Entire home/apt	89	1	194	270
5022	Manhattan	Entire home/apt	80	10	0	9
5099	Manhattan	Entire home/apt	200	3	129	74

Section 2: Supplementary Methods

The graph demonstrates the price level based on the room type. It was not readily interpretable due to too many outliers that the box plots were not observable. Therefore, it represents why the outliers should be removed for this analysis. [13]



Similar to the previous graph, the graphs below are the result when the outliers were not removed. They are looking hard to interpret due to the expansive price level.


Minimum Rental Days of Airbnb in NYC in Each Regions
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Price
Minimum Rental Days of Airbnb in NYC