Research on the Key Elements of Airbnb Customers' Room Monthly Reviews

Zongsheng Su [1, *, †]

[1]School of Professional Studies, New York University, New York, United States

*Corresponding author: nanafeifei0229@163.com

Abstract Airbnb has become a frequent option when people start to plan their trips. The trend of Airbnb usage has shown that people are switching from regular hotel booking platforms like Expedia, Bookings, etc., to Airbnb while traveling to a new place. Airbnb is not only a platform that offers new options for hotel booking but also a lifestyle-sharing platform that allows travelers to share their living experiences from different destinations. What's more interesting is that even during the Covid-19 pandemic, when people cannot travel or go outside, they can still interact with homeowners (hosts) on their chosen topic via streaming video. Therefore, those contents successfully attract more users who become passionate about the Airbnb experience and are willing to share their comments or reviews under each topic or location. This paper focuses on the renting industry. We select Airbnb and try to find the relationships behind different variables. We want to look at our database and use different languages to understand Airbnb in New York, such as multiple regression analysis. We have a database before the Covid-19 outbreak, which can fairly reflect the situation during the normal time. In order to find the relatively accurate correlation results, we want to confirm the correlations which confidently proclaimed that 'calculated_host_listings_count' has positively correlated with 'Reviews_per_month'. This paper aims to identify the key elements that most impact customers submitting their monthly reviews and how those reviews motivate other customers to book their upcoming trips through Airbnb continuously.

Keywords: Online Travel Booking, Reviews, Covid-19, Data Science, New York, Hotel industry.

1. Introduction

1.1 Research background

Every semester, at least three to five thousand Chinese students will choose to study abroad in NY, and most of them rent rooms in New York. Furthermore, many travelers prefer to use Airbnb as their first online travel agency rather than make price comparisons via multi-platforms. Therefore, we are interested in the renting industry and want to see how Airbnb users' behaviors make this platform more attractive than its competitors. Since Airbnb is the current industry leader with greater customer loyalty, why is Airbnb maintaining ongoing solid relationships is what we want to know. We hope to use specific models to determine the secrets to achieving customer loyalty and increasing customer retention with a higher conversion rate. We have looked at some statistics results of Airbnb, there are more than 12 million users in 2021, and 3 million users are coming from the Asia Pacific. Therefore, we select Airbnb and choose several types of variables to measure the behind relationships.

1.2 Research Objective

'Sharing economy' is no longer ideal for most industries, especially the hotel industry. Airbnb turns to be the major player favor in both the hotel industry and sharing economy in 2007. By the summer of 2016, more than 100 million guests had used Airbnb, and the service boasted over two million global listings [1]. On the other side, Airbnb poses a significant challenge to the traditional hospitality industry, since Lauritzen stated that the higher the price, the fewer reviews (-12%) and reviews per month (-22%) a host receives. The rationale might be a lower monthly booking rate for more costly residential properties. The less frequently a host is booked, the less feedback she receives[2]. This analysis does not include the time, price per person component, etc. we have to choose a narrow and deep assortment strategy with fewer variables to figure out which one or two variables are bringing more monthly reviews.

1.3 Research framework

1. How many factors impact Airbnb's guests' decision-making? 2. Why do people choose Airbnb instead of other online booking agencies? 3. How do different category factors impact customer reviews? Connecting with Harrison and Wicks's theory, this report supports the conflict between different stakeholders, and several correlational studies investigate relationships between provided variables [3].

2. Literature review

Tussyadiah and Zach proposed that peer-to-peer (P2P) is an emerging model in the tourism industry that has changed and disrupted the way consumers choose accommodation. P2P is a component of a bigger movement known as the "Sharing Economic," where consumers share and offer underutilized assets to other consumers [4]. Our study is different from Dann et al.'s research which focuses more on the reasons for guests‘ motivations to make decisions like trustworthiness [5]. We analyzed the different variables' impacts on the monthly ratings, including host, listing price, guest staying length, number of listings and locations, etc. According to Sherwood's report, Airbnb guests can book more than 6 million rooms, flats, and houses, in around 81.000 cities worldwide. More than 150 million hosts enrolled in the system, which provided accommodation services in 2018 for 260 million guests [6]. Meijerink and Schoenmakers found that customers' attitudes of satisfaction and loyalty are favorably and linearly associated with submitting customer reviews of an Airbnb stay. In other words, pleased customers are more likely than disgruntled customers to publish a review following an Airbnb stay[7].

3. Hypothesis development

3.1 Hypothesis testing

Null hypothesis: Variables that most affect the review per month are price, minimum nights, and the number of reviews.

Alternative hypothesis: Reviews per month may be affected by other factors.

$$y = \beta(0) + \beta1(x1) + \beta2(x2) + \beta3(x3) + \cdots + \varepsilon i \quad (1)$$

Hypothesis testing:

$$\text{Null (H0): } \beta(0) = \beta(1) = \beta(2) = \cdots \beta(k) = 0 \quad (2)$$

Alternative: Not H0

Null hypothesis state that all K covariates together do not explain a significant amount of the variation of y.

$$\text{Under the null hypothesis, the full model: } y = \beta0 + \varepsilon i \quad (3)$$

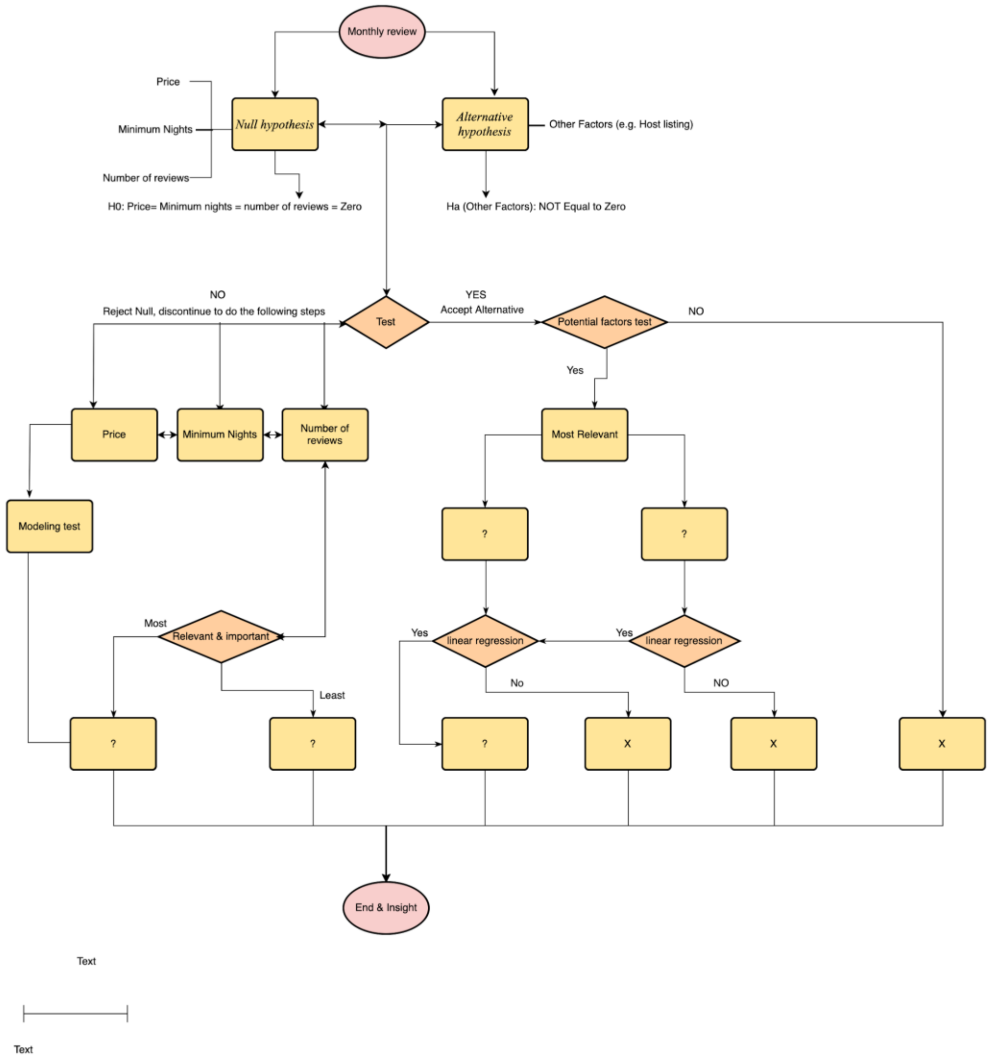

Figure 1. Process map

4. Methods

This study collected data from multi-source such as the Airbnb website, online booking agencies like Tripadvisor reviews, New York listings, and metrics open data from Kaggle, New York State Attorney General report, etc. Since we did not know exactly how many variables we needed and which variable may significantly affect the final results, we did data clean first and used regression analysis to see the relations that bring fewer variables to the further analysis.

Database maintenance: missing values, remove the availability=0

Census data: take a look at the mean and median values of variables. For example, the median price and mean price.

Enhancement data: we will create a new column to present the date from 7/8/2019, which is the latest update of this database.

RFM analysis: We would like to know how the customer performs based on our recode rules; Build multiple linear regression to find the relationship between variables with stepwise regression.

LTV analysis: Long-term rent will be lower profit for both Airbnb and hosts, but it will benefit the users. Short-term rent can give additional travel experience compared with hotel companies. It is about shopping behavior changing and branding consistency.

Future Monitoring Promotional Intensity analysis: 4Ps analysis requires new business models combining tech & promotion to increase Customer Stickiness which can also support the LTV.

R-studio: The regression analysis has been done with R-studio. Our goal is to see the trend and correlation between price, location, hosts, etc.

5. Results and Discussion

We found multiple relationships by reviewing a large amount of data and defining it with sort keys (sorting algorithms), but most of them are not quite relevant to the results we need. Therefore, we decided to create the model by ourselves, which first shows the significant variables, then represents the helpful dependent results with fewer samples and the right p-value. We need data maintenance first and test different variables' correlations instead of finding the key variables. We also want to prove the thought of Dana and David is True; they mentioned, 'In such low power experiments, p-value corrections are bound to be useful but blunt [8].' On the other hand, We mainly would like to determine what elements will affect the customers writing monthly reviews. We will also use different tools and strategies to find out if there is any insight behind Airbnb.

5.1 Data analysis

Data maintenance: We first decided to do database maintenance to support our analysis. We randomly select 1000 data rows from the 40 thousand rows' original database. *availability_365days* = 0 does not make sense because we have no idea when the room is unavailable, which will increase the bias. Therefore, we use SQL to remove it and remove missing value rows. Finally, we got 852 rows of data samples that we would like to use to analyze further.
SQL code: [select * from work.airbnb where availability_365days <> 0 quit;]

5.1.1 Variables we thought maybe needed

Id: room id;
Name: room description;
host_id: id for room host;
room_type: type of room;
Price: every rooms price for each night;
min_nights: the minimum night for the room that the customer should reserve;
#_of_reviews: the total number of reviews from customers who have lived in the room and written a review for the room.
Last_review: the date of the newest customer's review;
Reviews_per_month: average number of reviews for each month;
calculateed_host_listings_count: number of same room;
availability_365days: the number of days available in a year.
days from the latest day: calculatedulate the number of days from the last review date which transfers the date columns to statistically meaningful numbers
if private: we would like to analyze the room type, and we think only Chi-square is not enough. Therefore, we would like to transfer it to an ordinal number.

5.1.2 Nominal data

Before we have the linear regression, we use the Chai square test to detect if the nominal variable *room_type* influences the monthly review. We separate the monthly review number to be high and low, depending on whether the number is higher than the avg monthly review (=1.8). We get the result from the R-studio showed no significant difference between room types on monthly review. It seems people who would like to write the review have nothing to do with room type.

5.1.3 Question

Detect if the nominal variable room_type influences the monthly review.
Result:

No significant difference between room types on the monthly review.

Table 1. Summary of Airbnb reviews per month

```
> summary(airbnb$reviews_per_month)
```

Min.	1st Qu.	Med	Mean	3rd Qu.	Max.
0.020	0.490	1.230	1.818	2.750	11.400

Values	Column Labels 0	1	Grand Total
Count of If private	428	423	851
Avg of price	208.364486	102.2033097	155.5957697
Avg of min _nights	9.721962617	5.48463357	7.615746181
Avg of #_of _reviews	37.22897196	38.9787234	38.0987074
Avg of reviews_per_month	1.83521028	1.80106383	1.818237368
Avg of days from the lastest day	102.3084112	103.8676123	103.0834313
Max of days from the lastest day	1868	2479	2479
Min of days from the lastest day	0	1	0
Avg of availability_365days	163.635514	185.9361702	174.720329

Table 2. Chi-squared Test

```
> View(airbnb)
> table(airbnb$room_type,airbnb$' high/low')

                 high low
Entire home/apt 158 250
Private room          166 257
Shared room               6    14
> chisq.test(airbnb$room_type,airbnb$'high/low',correct=FALSE)

          Pearson's Chi-squared test

data: airbnb$room_type and airbnb$'high/low'
X-squared = 0.68821, df = 2, p-value = 0.7089

> table(airbnb$'If private',airbnb$'high/low')

   high low
0    164 264
1    166 257
> chisq. test(airbnb$'If private',airbnb$'high/low', correct=FALSE)

        Pearson's Chi-squared test

data: airbnb$'If private'and airbnb$'high/low'
X-squared = 0.076797, df = 1, p-value = 0.7817
```

5.2 Linear Regression Model

We insert all ratio variables and use R-studio to run a linear regression. Besides, we set 95% as our confidence interval, and the p-value < 0.05 is the significant variable. According to the result, we found that *#_of_reviews*, *calculated_host_listings_count*, and *days from the lastest day* are significant variables and impact monthly reviews.

Table 3. Significant variable Test

Call:

lm(formula = airbnb$reviews_per_month ~ airbnb$price + airbnb$min_nights + airbnb$#_of_reviews + airbnb$calculated_host_listings_count + airbnb$availability_365days + airbnb$ `days from lastest day'

Residuals:

Min	1Q	Med	3Q	Max
-2.4054	-1.1464	-0.5184	0.8223	9.3301

C.:

	Estimate	Std. Error	
(Intercept)	2.2329132	0.1099202	
airbnb$price	-0.0001660	0.0002748	
airbnb$min_nights	-0.0037449	0.0033671	
airbnb$#_of_reviews	-0.0027072	0.0010107	
airbnb$calculated_host_listings_count	0.0045009	0.0014587	
airbnb$availability_365days	-0.0003426	0.0004455	
airbnb$'days from lastest day’	-0.0023285	0.0002354	
	t value	Pr(>ltl)	
(Intercept)	20.314	< 2e-16	***
airbnb$price	-0.604	0.54590	
airbnb$min_nights	-1.112	0.26637	
airbnb$#_of_reviews	-2.678	0.00754	**
airbnb$calculated_host_listings_count	3.086	0.00210	**
airbnb$availability_365days	-0.769	0.44215	
airbnb$'days from lastest day’	-9.890	< 2e-16	***

Significant. codes: 0 '***’ 0.001 ‘**’ 0.01 '*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual Std. error: 1.567 on 844 degrees of freedom
Multiple R-squared: 0.1388, Adjusted R-squared: 0.1327
F-statistic: 22.67 on 6 and 844 DF, p-value: < 2.2e-16

5.3 Multiple Regression Model

We use R-studio to run the Anova test and get the *#_of_reviews*, *calculated_host_listings_count*, *days from the lastest day, and availability_365days* as our statistically Significanticant variables. There is one variable different from a simple linear regression model.

Table 4. Anova Test

> anova(fit)
Analysis of Variance Table

Response: airbnb$reviews_per_month

	Df	Sum Sq	Mean Sq	F value
airbnb$price	1	1.22	1.219	0.4966
airbnb$min_nights	1	0.00	0.000	0.0002
airbnb$#_of_reviews	1	51.62	51.624	21.0337
airbnb$calculated_host_listings_count	1	29.48	29.481	12.0116
airbnb$availability_365days	1	11.50	11.505	4.6875
airbnb$'days from lastest day'	1	240.07	240.067	97.8121
Residuals	844	2071.49	2.454	

	Pr(>F)	
airbnb$price	0.4812099	
airbnb$min_nights	0.9897957	
airbnb$#_of_reviews	5.196E-06	***
airbnb$calculated_host_listings_count	0.0005554	***
airbnb$availability_365days	0.0306623	*
airbnb$'days from lastest day'	< 2.2e-16	***
Residuals		

Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

5.4 Correlation Matrix

Then we use R-studio to look at the correlation of all the variables. All the correlation results are small, and there is no multicollinearity in the model.

Table 5. Correlation of all the variables

> subdata<-c("price", "min_nights ", "#_of_ reviews" , " reviews_per_month" , "calculated_ host_ listings_count ","availability_ 365days", "days from lastest day")

> sub<-airbnb[subdata]

> cor(sub,method="pearson")

	price	min_nights	#_of_ reviews	reviews_per_ month	calculated_ host_ listings_count	availability_ 365days	days from lastest day
price	1.000000000	-0.0090358554	-0.08519004	-0.0225093864	0 .08465104	0.09094039	0.05276361

min_nights	-0.009035855	1.0000000000	-0.16668589	0.0006120335	0.14161897	0.10676344	-0.03282550
#_of_reviews	-0.085190043	-0.1666858874	1.00000000	-0.1420410316	-0.11880522	0.02446285	0.15040652
reviews_per_month	-0.022509386	0.0006120335	-0.14204103	1.0000000000	0.11997054	-0.06873580	-0.34597072
calculated_host_listings_count	0.084651037	0.1416189725	-0.11880522	0.1199705369	1.00000000	0.09749197	-0.05528179
availability_365days	0.090940391	0.1067634409	0.02446285	-0.0687358036	0.99749197	1.00000000	0.14072297
days from lastest day	0.052763610	-0.0328255009	0.15040652	-0.3459707195	-0.05528179	0.14072297	1.00000000

5.5 Backward and Forward Regression

We did both backward and forward regression and got the same results. However, the result of backward regression in our case is more precise. Therefore, we have our final model.
Monthly reviews=2.118-0.0025 number of reviews + 0.004 calculated host listings account – 0.0024 days from the latest day.

5.6 Explanations

Based on Hati, Balqiah, Hananto, and Yuliati's review, most articles on Airbnb published in peer-reviewed journals discuss guests or customers of Airbnb (93 articles; 32%), followed by hosts (88 articles; 30%), communities (56 articles; 19%), competitors (32 articles; 11%), government (24 articles; 8%), and employees (0 articles; 0%) [9]. All relationships can be logically understood. If the total number of reviews increases, there is a decrease in monthly reviews. It is probably because Airbnb already has enough reviews, and customers have an overview of the room. They do not have surprises in the room or different opinions from prior reviews. Therefore, with the total number of reviews increasing, people do not want to write more comments. The monthly comments will decrease when an extra total number of reviews increases with a decreased rate.

Calculated host listing accounts have a positive relationship with monthly reviews. Usually, people write comments because they have very good or bad experiences. Therefore, the positive relationship may be because the more rooms a host owns, the more experience they have. The hosts understand the importance of reviews and have characteristics like a good living experience, better room management, or an encouraging review process.

Days from the latest day have a negative relationship with monthly reviews because the more recent reviews added will increase the average monthly reviews.

Table 6. Backward Regression

>summary(step(fit,direction-"backward"))

>Start: AIC=771.06

airbnb$reviews_per_month ~ airbnb$price + airbnb$min_nights + airbnb$#_of_reviews + airbnb$calculated_host_listings_count +airbnb$availability_365days + airbnb$ `days from lastest day'

	Df	Sum of Sq	Rss
- airbnb$price	1	0.896	2072.4
- airbnb$availability_365days	1	1.451	2072.9
- airbnb$min_nights	1	3.036	2074.5
<none>			2071.5
- airbnb$#_of_reviews	1	17.608	2089.1
- airbnb$calculated_host_listings_count	1	23.368	2094.9
- airbnb$'days from lastest day'	1	240.067	2311.6

	AIC
- airbnb$price	769.43
- airbnbSavailability_365days	769.65
- airbnb$min_nights	770.31
<none>	771.06
- airbnb$#_of_reviews	776.26
- airbnb$calculated_host_listings_count	778.61
- airbnb$'days from lastest day'	862.37

Step: AIC=769.43

airbnb$reviews_per_month ~ airbnb$min_nights + airbnb$#_of_reviews + airbnb$calculated_host_listings_count + airbnb$availability_365days + airbnb$ `days from lastest day'

	Df	Sum of Sq	RSS
~ airbnb$availability_365days	1	1.654	2074.0
~ airbnb$min_nights	1	2.906	2075.3
<none>			2072.4
~ airbnb$#_of_reviews	1	17.029	2089.4
~ airbnb$calculated_host_listings_count	1	22.822	2095.2
~ airbnb$'days from latest day’	1	242.554	2314.9

	AIC
~ airbnb$availability_365days	768.11
~ airbnb$min_nights	768.62
<none>	769.43
~ airbnb$#_of_reviews	774.39
~ airbnb$calculated_host_listings_count	776.75
~ airbnb$'days from latest day’	861.62

Table 6.1. Backward Regression

>Step: AIC=768.11

airbnb$reviews_per_month ~ airbnb$min_nights + airbnb$#_of_reviews + airbnb$calculated_host_listings_count + airbnb$`days from lastest day'

	Df	Sum of Sq	Rss
- airbnb$min_nights	1	3.409	2077.4
<none>			2074.0
- airbnb$#_of_reviews	1	17.375	2091.4
- airbnb$calculated_host_listings_count	1	21.870	2095.9
- airbnb$'days from lastest day'	1	253.664	2327.7

	AIC
- airbnb$min_nights	767.50
<none>	768.11
- airbnb$#_of_reviews	773.21
- airbnb$calculated_host_listings_count	775.03
- airbnb$'days from lastest day'	864.30

Step: AIC=767.5

airbnb$reviews_per_month ~ airbnb$#_of_reviews + airbnb$calculated_host_listings_count + airbnb$'days from lastest day'

	Df	Sum of Sq	RSS
<none>			2077.4
~ airbnb$#_of_reviews	1	15.487	2092.9
~ airbnb$calculated_host_listings_count	1	20.087	2097.5
~ airbnb$'days from latest day’	1	253.476	2330.9

	AIC
<none>	767.50
~ airbnb$#_of_reviews	771.82
~ airbnb$calculated_host_listings_count	773.69
~ airbnb$'days from latest day’	863.47

Call:

lm(formula = airbnb$reviews_per_month ~ airbnb$#_of_reviews + airbnb$calculated_host_listings_count + airbnb$'days from lastest day')

Residuals:

Min	1Q	Med	3Q	Max
-2.335	-1.1387	-0.5052	0.8345	9.3497

C.:

Table 6.2. Backward Regression

C.:

	Estimate	Std. Error		
(Intercept)	2.1180631	0.0703846		
airbnb$#_of_reviews	-0.0024981	0.0009942		
airbnb$calculated_host_listings_count	0.0041108	0.0014364		
airbnb$'days from lastest day'	-0.0023636	0.0002325		
	t value		Pr(>ltl)	
(Intercept)	30.093	<	2e-16	***
airbnb$#_of_reviews	-2.513		0.01216	*
airbnb$calculated_host_listings_count	2.862		0.00432	**
airbnb$'days from lastest day'	-10.166	<	2e-16	***

Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual Std. error: 1.566 on 847 degrees of freedom
Multiple R-squared: 0.1363, Adjusted R-squared: 0.1333
F-statistic: 44.57 on 3 and 847 DF, p-value: < 2.2e-16

Table 7. Forward Regression

```
>summary(step(fit,direction-"forward"))
>Start:   AIC=771.06
airbnb$reviews_per_month ~ airbnb$price + airbnb$min_nights +
airbnb$#_of_reviews + airbnb$calculated_host_listings_count +
airbnb$availability_365days + airbnb$ `days from lastest day'

Call:
lm(formula = airbnb$reviews_per_month ~ airbnb$price + airbnb$min_nights +
airbnb$#_of_reviews + airbnb$calculated_host_listings_count +
airbnb$availability_365days + airbnb$ `days from lastest day'
```

Residuals:

Min	1Q	Med	3Q	Max
-2.4054	-1.1464	-0.5184	0.8223	9.3301

C.:

	Estimate	Std. Error	t value	Pr(>ltl)	
(Intercept)	2.2329132	0.1099202	20.314	< 2e-16	***
airbnb$price	-0.0001660	0.0002748	-0.604	0.54590	
airbnb$min_nights	-0.0037449	0.0033671	-1.112	0.26637	

airbnb$#_of_reviews	-0.0027072	0.0010107	-2.678	0.00754	**
airbnb$calculated_host_listings_count	0.0045009	0.0014587	3.086	0.00210	**
airbnb$availability_365days	-0.0003426	0.0004455	-0.769	0.44215	
airbnb$'days from lastest day'	-0.0023285	0.0002354	-9.890	< 2e-16	***

Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual Std. error: 1.567 on 844 degrees of freedom
Multiple R-squared: 0.1388, Adjusted R-squared: 0.1327
F-statistic: 22.67 on 6 and 844 DF, p-value: < 2.2e-16

5.6.1 Between Variables

We also take a look at the relationship of price with other variables to get a deeper insight.

●Y =β_0+β_1 x_1+β_2 x_2+β_3 x_3+β_4 x_4+β_5 x_5 (1)

●Price as dependent variable to several independent variables x

●β_0,1,2,3,4,5 – coefficients(C.) (2)

●x_1- Min Nights, x_2 -# of reviews , x_3- Reviews per moths, x_4-Calculated host listing count, x_5- Availability in 365days days

●The initial model is Price = 152.03735 - 0.53552MinNights - 0.32832#ofReviews - 4.53720 ReviewsPerMonth + 0.40030 CalculatedHostListingsCount + 0.14078 Availabilityin365days (3)

Table 8. Coefficients result

C.:

	Estimate	Std.Error	t value	Pr(>ltl)	
(Intercept)	152.03735	15.49103	9.815	< 2e-16	***
MinNights	-0.53552	0.42172	-1.270	0.20449	
#ofReviews	-0.32832	0.12601	-2.606	0.00933	**
ReviewsPerMonth	-4.53720	4.08055	-1.112	0.26649	
CalculatedHostListingsCount	0.40030	0.18328	2.184	0.02923	*
Availabilityin365days	0.14078	0.05522	2.550	0.01096	*

●Backward Elimination Procedure comparing the AIC value

●Lowest AIC=8990.98

The Final Model: Price = 138.96810 - 0.28585# of Reviews + 0.35109CalculatedHostListingsCount + 0.13843Availabilityin365days (4)

Table 9. Final Model

```
> fit<-lm(airbnb$calculated_host_listings_count~airbnb$availability_365days)
> summary(fit)
```

Call:
lm(formula = airbnb$calculated_host_listings_count ~ airbnb$availability_365days)

Residuals:

Min	1Q	Med	3Q	Max
-14.14	-9.70	-5.94	-3.65	320.30

C.:

	Estimate	Std. Error	t value	Pr(>ltl)	
(Intercept)	4.29299	2.22881	1.926	0.05442	.
airbnb$availability_365days	0.02973	0.01042	2.854	0.00442	**

Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual Std. error: 37.53 on 849 degrees of freedom
Multiple R-squared: 0.009505, Adjusted R-squared: 0.008338
F-statistic: 8.147 on 1 and 849 DF, p-value: 0.004418

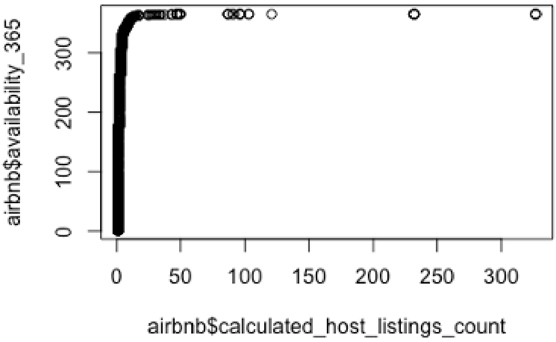

Figure 2. Relationship BW Variables

6. Insights about Price

6.1 Supporting Strategy

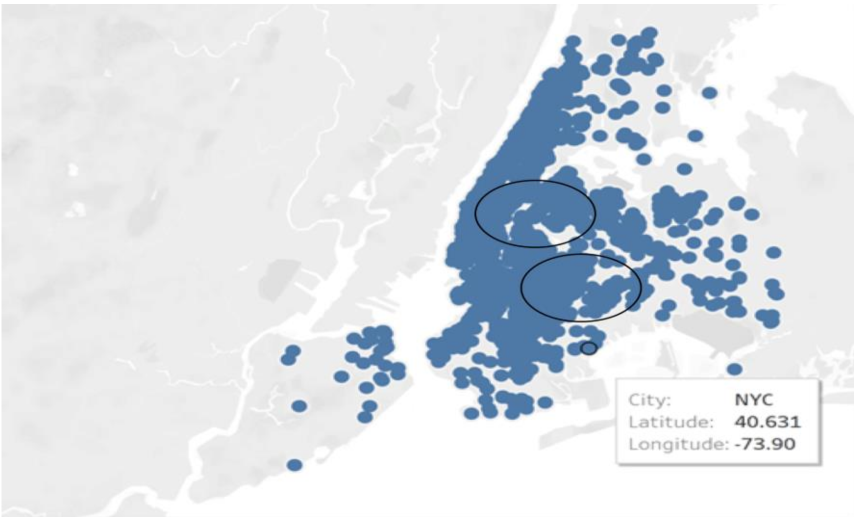

Figure 3. New York City Map

NYC Airbnb Listings Distribution Status: Most Airbnb listings in the New York area are mainly distributed in the Manhattan and Brooklyn areas, which are more tourist attractions and higher demand for renters who need lower rent. As stated in Gunter's research, 'More wealthy Airbnb users are also willing to pay premium rates for Airbnb listings that are roomy and/or operating in Manhattan, i.e., close to New York City's major attractions[10].'

6.2 Methodology (Same as 'Between Variables' section)

●y =β0+β1x1+β2x2+β3x3+β4x4+β5x5 (1)
●Price as the dependent variable to several independent variables x
●β0,1,2,3,4,5 – coefficients(C.) (2)
●x1- Min Nights, x2 - # of reviews, x3 - Reviews per moths, x4 - Calculated host listing count, x5 - Availability in 365 days
●The initial model is：
Price = 152.03735 - 0.53552MinNights - 0.32832 #ofReviews - 4.53720 ReviewsPerMonth + 0.40030 CalculatedHostListingsCount + 0.14078 Availabilityin 365 days (3)

6.3 Expected Profit Calculations

We assume the room price is an average room price of $155 and spend $46.5 for paid. Total revenue will be $108.5 for each pay. And we cost $2 as a promotion fee on Airbnb. The cost of an unpaid order is $1.

Table 10. Ben and Jones Expected Profit

Customers	Probability of ordering	Probability of not paying
Ben	10%	3%
Jones	20%	5%

Ben = 0.1*0.97*(108.5-2)+0.1*0.03*(-2-1)+0.9*(-2)=10.33-0.09-1.8=8.44
Jones = 0.2*0.95*106.5+0.2*0.05*(-3) + 0.8*(-2)=20.235-0.03-1.6=18.605

6.3.1 Lifetime value

We set annual interest rate = 2%, pmt per month= price*monthly review, and calculated the present value of the future one year for each room.

Table 11. Lifetime Value

= -PV(0.02/12,12,A23*E23)									
A	B	C	D	E	F	G	H	I	J
price	min_nights	#_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365days	latest day	days from last est day	LTV
125	3	105	5/27/19	1.41	1	195	7/8/19	42	$ 2,092.26
200	2	109	3/14/18	0.25	1	336	7/8/19	481	$ 593.55
58	1	104	6/24/19	3.7	9	300	7/8/19	14	$ 2,547.52
325	2	0	6/18/19	4.12	3	268	7/8/19	20	$ 15,895.28
100	1	12	7/7/19	4.25	1	89	7/8/19	1	$ 5,045.18

56	1	20	2/15/19	0.21	4	129	7/8/19	143	$ 139.60
50	1	84	5/8/19	1.21	4	302	7/8/19	61	$ 718.20
50	3	15	2/14/19	3.04	2	217	7/8/19	144	$ 1,804.39
150	3	154	3/8/17	0.89	1	218	7/8/19	852	$ 1,584.78
395	5	4	6/20/19	1	1	167	7/8/19	18	$ 4,689.05
200	2	0	6/7/19	2.68	4	189	7/8/19	31	$ 6,362.86
250	3	37	1/1/19	0.18	2	140	7/8/19	188	$ 534.20
69	2	166	1/4/19	0.42	2	140	7/8/19	185	$ 344.02
120	1	134	5/5/19	0.57	6	249	7/8/19	64	$ 811.98
75	2	2	7/1/19	0.98	1	188	7/8/19	7	$ 872.52
649	2	13	12/6/16	0.03	12	365	7/8/19	944	$ 231.13
69	4	5	9/2/16	0.53	4	98	7/8/19	1039	$ 434.12
80	1	12	6/23/19	0.4	1	59	7/8/19	15	$ 379.87
110	1	21	7/1/19	0.88	12	256	7/8/19	7	$ 1,149.11
109	30	43	6/26/19	1	2	365	7/8/19	12	$ 1,293.94
189	30	0	6/16/19	4.52	87	365	7/8/19	22	$ 10,141.16
3000	3	0	6/2/19	1.88	1	23	7/8/19	36	= -PV(0.02/12,12,A23*E23)
95	2	4	6/30/19	1.7	2	28	7/8/19	8	$ 1,917.17
158	2	20	6/3/19	0.57	1	2	7/8/19	35	$ 1,069.10
300	2	62	5/23/19	0.9	1	274	7/8/19	46	$ 3,205.17
184	3	146	5/26/19	0.27	1	143	7/8/19	43	$ 589.75
80	3	10	6/18/19	3.13	1	332	7/8/19	20	$ 2,972.50

6.3.2 CHAID analysis - Decision Tree

Using Rpart and Rpart. Plot packages, we built a decision tree as below.

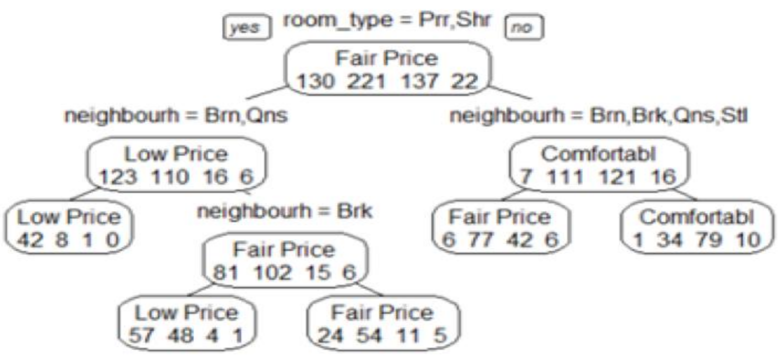

Figure 4. Decision Tree

6.3.3 Train set and Validation set:

Table 12. Train Test

Confusion Matrix and statistics

Prediction	Reference Low Price	Fair Price	Comfortable Price	Luxury Price
Low Price	99	56	5	1
Fair Price	30	131	53	11
Comfortable Price	1	34	79	10
Luxury Price	0	0	0	0

Overall Statistics

Accuracy:	0.6059
95% CI:	(0.562, 0.6486)
No information Rate:	0.4333
P-Value (Acc > NIR):	3.716e-15
Kappa:	0.4056

Table 13. Validation Test

Confusion Matrix and statistics

Prediction	Reference Low Price	Fair Price	Comfortable Price	Luxury Price
Low Price	67	39	3	1
Fair Price	19	88	42	4
Comfortable Price	0	27	47	4
Luxury Price	0	0	0	0

Overall Statistics

Accuracy:	0.5924
95% CI:	(0.5381, 0.645)
No information Rate:	0.4516
P-Value (Acc > NIR):	1.26e-7
Kappa:	0.377

The Accuracy of the train set is 0.6059, and the accuracy of the validation set is 0.5924 which is almost the same as the train set. This result means our model is good to use.

7. Conclusion

7.1 Implication

Overall, we found that the number of reviews, calculated host listings accounts, and days from the latest day have a statistically significant impact on the monthly reviews. Therefore, the null hypothesis is rejected, and the new equation is accepted. Besides, when specifically focusing on the room price, it demonstrates that the number of reviews, calculated host listings count, and availability in 365 days will statistically significantly affect the room price. However, there is a negative relationship between price and the number of reviews. Part of the reason might be the high room price has fewer customers and fewer reviews. The three variables, including calculated host listings, availability in 365 days, and price, are significantly related. Usually, hosts with more rooms will treat Airbnb as a business. They tend to pay more attention to room decoration and management. Thus, the room will provide a better customer experience. These hosts will spend more money than other hosts who only lend their homes several days a year. It makes sense that availability and price increase with calculated host listings. It also can be explained that higher calculated host listings will bring a better experience and more monthly reviews.

We use MAP, Expected profit calculation, Lifetime value, and Chaid Analysis as our supporting strategy to help better understand the database and future marketing plans. The host with more rooms will cost more than other hosts who only lend their home several days a year. It makes sense that availability and price are increased with calculated host listings. It also can be explained that higher calculated host listings will bring a better experience and more monthly reviews.

7.2 Limitations and Future studies

This article lacks primary data and mainly uses secondary data, so it may contain certain biases. It is possible to face a problem that the more complex the model becomes, the higher the distortion of simulation results may cause.

In the future, primary data can be obtained through surveys, interviews, etc. The survey settings mainly start from the user experience, focusing on usage habits, portability, problems, and potential needs in the follow-up process.

Acknowledgments

First and foremost, I am grateful to Prof. Amreeta Choudhury, Prof. Jeremy Cohn, and Prof. Richard L Gant of New York University for their tremendous counsel, persistent support, and patience during my M.A. studies. Their vast knowledge and broad experience have inspired me throughout my academic study and daily life.

Nobody has been more essential to me in the pursuit of this project than my family members. I'd want to thank my parents for their love and guidance in anything I do. They are the greatest role models.

Last, my friends (Nico) Peiying Shen, (Elva) Yuanji Li, Lok Cheung, (Richard) Ming Mu, (Catherine) Xinrui Zuo, (Sonic) Qi Jin and (Laurinda)Tingyuan Luo, have always been at my side. I could not have done this work without their assistance and consultation.

References

[1] M. Chafkin, and E. Newcomer. "Airbnb Faces Growing Pains as It Passes 100 Million Users." Bloomberg, 11 July 11. 2016. http://www.bloomberg.com/news/articles/2016-07-11/airbnbfaces-growing-pains-as-it-passes-100-million-users.

[2] K. Lauritzen, "How to Maximize Profits on Airbnb? Data-Based Approach for Hosts." Medium, Towards Data Science, 7 Nov. 2019. https://towardsdatascience.com/how-to-maximize-profits-on-airbnb-data-based-approach-for-hosts-beaf08f26941.

[3] Harrison JS, Wicks AC. Stakeholder theory, value, and firm performance. Business ethics quarterly. 23 Jan. 2013; (1): 97-124. https://www.cambridge.org/core/journals/business-ethics-quarterly/article/abs/stakeholder-theory-value-and-firm-performance/B84D07B5B6241F83EC7D904E28B21015

[4] Tussyadiah IP, Zach F. Identifying salient attributes of peer-to-peer accommodation experience. Journal of Travel & Tourism Marketing. 13 Jun. 2017; 34(5):636-52. https://www.tandfonline.com/doi/full/10.1080/10548408.2016.1209153?casa_token=FOUgf2PxiwoAA

AAA%3ALJMA1IDRcgg1AeijdTxQbBRWrU4VeNcwomR_K1TEQmXyUi5jqIa9puVqT9aib3meyX81cK4FeCsUUQ

[5] Dann D, Teubner T, Weinhardt C. Poster child and guinea pig–insights from a structured literature review on Airbnb. International Journal of Contemporary Hospitality Management. 17 Dec. 2018. https://www.emerald.com/insight/content/doi/10.1108/IJCHM-03-2018-0186/full/html?casa_token=wNgepI6jUCMAAAAA:CHmWPpOqiWvvHeDqsofVN70eW2ii-RicXzTOdB2vjTfxrN8U6LaI5vRjprpTuyBz66a4zHPJRF2cnil7APPhd05j9FjdMSZN6WWZiFALPAljwX6Z_bE_

[6] S. Harriet. “How Airbnb took over the world.” The Guardian, 5 May 2019. https://www.theguardian.com/technology/2019/may/05/airbnb-homelessness-renting-housing-accommodation-social-policy-cities-travel-leisure

[7] Meijerink, J. and Schoenmakers, E., “Why are online reviews in the sharing economy skewed toward positive ratings? Linking customer perceptions of service quality to leaving a review of an Airbnb stay”, Journal of Tourism Futures, 2021, Vol. 7 No. 1, pp. 5-19. https://doi.org/10.1108/JTF-04-2019-0039

[8] Pascovici D, Handler DC, Wu JX, Haynes PA. Multiple testing corrections in quantitative proteomics: A useful but blunt tool. Proteomics. Sep 2016;16(18):2448-53. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/pmic.201600044?casa_token=0I3xSbvqs34AAAAA%3AHbZjJZYiCEQQyXeJhdbJs3HatKLpxihy6tWy6dr7HKaFoL1sfhc9mSgzXliGGpgxhOXUoG7pLlAs_3PB

[9] Hati SR, Balqiah TE, Hananto A, Yuliati E. A decade of systematic literature review on Airbnb: the sharing economy from a multiple stakeholder perspective. Heliyon. 1 Oct. 2021;7(10):e08222. https://www.sciencedirect.com/science/article/pii/S2405844021023252

[10] Gunter U, Önder I, Zekan B. Modeling Airbnb demand to New York City while employing spatial panel data at the listing level. Tourism Management. 1 Apr. 2020; 77:104000.https://www.sciencedirect.com/science/article/pii/S0261517719301980?casa_token=6lDclKXCnyIAAAAA:Jbi1sauiHJpfpeQGg_fFX-SCtWSqSYy7uuYlSe5vi9gt_Eu_P2HHe-ymTObp7YEPxSdPVnJP