Pricing Analytics of the Sharing Economy in Lodging—A Case Study of the Airbnb Online Marketplace

Yiming Peng[1,a,*]

[1]Beijing City International School, Beijing, China

[a]staciepeng0919@gmail.com

*Corresponding author

Keywords: Sharing Economy, House Rental, Airbnb, Big Data Analysis, Econometrics

Abstract. This paper studies the pricing strategy of Airbnb's online marketplace. Through the analysis of Airbnb's publicly available listing data over time, we use econometric methods including statistical regression and time series analysis to understand how the Airbnb listings' pricing is affected by their geofigureical locations, room-specific features (room type, number of beds, etc.), customer reviews and potential seasonality effects. We aim at gaining insights into Airbnb's pricing strategy and providing observations and guidance on the operations of the two-sided markets and more broadly, the proliferating sharing economy. This study applies econometrics and big data analysis techniques to a practical setting. We use the Python's Pandas package to conduct basic data analysis and visualization, and then we employ regression analysis and time series analysis to analyze the impact of various factors on pricing and make price predictions. In order to make the prediction more accurate, we use resampling method to get the average coefficient of the regression. Lastly, we draw conclusions from our analysis and provide practical insights and understanding of the Airbnb platform and the sharing economy it represents.

1. Introduction

1.1 Motivation and research questions

In recent years, sharing economy has flourished around the world, and the great popularity of smartphones and mobile internet accelerated the emergence and expansion of various forms, such as online car rental, shared bicycle rental, short-term house rental, and the gig economy. Airbnb is a great example of online marketplaces and the sharing economy. With the rise of Airbnb, the short-term house rentals have grown rapidly. The growth of short-term rentals helps tourists experience local life, understand local culture and create employment opportunities for local residents. Based on online statistics [1], Airbnb has over 150 million users worldwide, there are Airbnb homes in 81,000 cities in 2018, generating $93 million in profit in 2017, and Airbnb has 153% global compound growth rate since 2009. Hence, it is an important factor for the growth in sharing economy. Since Airbnb has significantly affected the development of house rental markets, people may wonder how Airbnb is being used in their daily life [2]. Questions such as "How many listings are in my neighbourhood and where are they?", "How many houses and apartments are being rented out frequently to tourists and not to long-term residents?", and "how much are hosts making from renting to tourists (compare that to long-term rentals)?" motivated our study in gaining insights about this market and how the house rental market brings huge economic benefits.

However, in the past, house hosts have relied heavily on qualitative research on the pricing of houses with studies of tourists and their behavior.[3] For example, they evaluate pricing strategies for housing quality and user satisfaction through users' feedback. These methods do not quantitatively reflect the behavior and preferences of individual users and provide the hosts with guidance on house rental price.

In this paper, we focus on the short-term house rental markets, the world's largest online platform for share house rental Airbnb to conduct big data analysis. Through econometric modeling, statistical regression, time series analysis, etc., we study the pricing strategy of Airbnb hosts, and analyze how the characteristic of the house (room, bed number, etc.), location, and users' review affect the pricing. We are able to provide guiding suggestions for both hosts and users.

In our research, we conduct the big data analysis based on three datasets from InsideAirbnb and one from the geofigureical information of New York City. Firstly, listing dataset (data_listing) is the summary of listings in Airbnb New York City and geofigureical information of these listings in June 2019. There are 48801 listings in NYC during June 2019 in total. Detailed information like host id, neighbourhood group, latitude, longitude, price, minimum nights and etc. were listed in columns. Secondly, dataset of listing details (data_listing_detail) contains detailed information of Airbnb listings in New York City during June 2019. There are 48801 rows and 106 columns in this dataset. Information including listing id, summary about the house, space description about the house, cancellation policy, guest requirements, reviews about the house, hosts' information, and etc. It offers more detailed information than the data_listing dataset. By calculating some basic statistics and visualizing these two data sets, we can have a better understanding on price in relation to other variables. Thirdly, Calendar dataset (data_calendar) contains information of Airbnb listings in New York City during June 2019 to June 2020 from insideairbnb.com. There are 17812547 rows and 7 columns in this data set. Total rows are calculated using 48801 (total number of listings) times 365/366 days. Information including listing id, dates, price, adjusted price, availability during specific dates, minimum nights and maximum nights. Lastly, in order to elaborate the analysis of Listing and Calendar with New York City geofigureical information, we import the geofigureical information dataset of New York City and combine it with the Airbnb listing data for a more comprehensive study of pricing. These datasets lay the foundation for subsequent econometric modeling: linear regression and time series analysis, and through these models we are able to give interesting insights for house rental market as well as future guidance for big data analysis of Sharing Economy.

1.2 Literature review

Airbnb, as an online house rental platform, is "a trusted community marketplace for people to list, discover, and book unique accommodations around the world" [4]. On this online marketplace, accommodation ranges from house, apartment, boutique hotel to even tree house. There are different room types such as private room, shared room and entire house serve for users.

The majority of research about Airbnb is based on quantitative methods (61.5%) and qualitative studies (18.5%) [5]. For quantitative methods, since Airbnb itself does not provide a detailed breakdown for all listings, a company fact sheet in 2012 stated that overall 57% of its listings were entire homes, 41% were private rooms, and 2% were shared spaces. Limited information is available regarding user demofigureics, but in 2012 Airbnb reported that approximately 40% of its guests were American, with Europeans comprising the majority of the rest [6]. For qualitative analysis, data publicly available on the Airbnb website (e.g. listing attributes or guest reviews) were most commonly used [5].

In our research, Pricing Analytics of the Sharing Economy in Lodging- A Case Study of the Airbnb Online Marketplace, different from the past studies, we employ big data analysis based on a wide-range and great amount of data (3 datasets with over 18 million rows) for many aspects and use quantitative method to analyze factors including the length of customer reviews. In addition, the object we study is a new and growing form of Sharing Economy compared to the existing research of online marketplaces. Therefore, this research will provide a more detailed quantitative analysis for Airbnb's pricing strategy at better accuracy.

2. Data analytics

2.1 Listing dataset

2.1.1 Data overview

Table 1, the listing data frame (data_listing) is the summary of New York City and geofigureical information in June 2019. There are 48801 listings in NYC during June 2019. Detailed information like host id, neighbourhood group, latitude, longitude, price, minimum nights, etc. listed in columns.

Table 1: Airbnb sample listings in NYC during June 2019

	2539	2595	3647	3831	4989
name	Clean & quiet apt home by the park	Skylit Midtown Castle	THE VILLAGE OF HARLEM....NEW YORK !	Cozy Entire Floor of Brownstone	Great 1 bdrm. apartment in the PERFECT location!
host_id	2787	2845	4632	4869	7118
host_name	John	Jennifer	Elisabeth	LisaRoxanne	New-Yorker
neighbourhood_group	Brooklyn	Manhattan	Manhattan	Brooklyn	Manhattan
neighbourhood	Kensington	Midtown	Harlem	Clinton Hill	Hell's Kitchen
latitude	40.64749	40.75362	40.80902	40.68514	40.7626
longitude	-73.97237	-73.98377	-73.9419	-73.95976	-73.99304
room_type	Private room	Entire home/apt	Private room	Entire home/apt	Entire home/apt
price	149	225	150	89	105
minimum_nights	1	1	3	1	4
number_of_reviews	9	44	0	258	27
last_review	2018/10/19	2019/5/7	NaN	2019/5/20	2018/7/25
reviews_per_month	0.21	0.38	NaN	4.53	0.24
calculated_host_listings_count	6	2	1	1	1
availability_365	365	331	365	182	83

For table 2, by grouping listings into different neighbourhood groups (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) and calculating the mean value of each columns, it suggests that houses in Manhattan the highest average price, houses in Brooklyn the lowest average availability in 365 days, and houses in Queens the highest average reviews per month.

Table 2: Listing grouped by neighbourhood group and mean values

neighbourhood_group	Bronx	Brooklyn	Manhattan	Queens	Staten Island
id	2.19E+07	1.77E+07	1.83E+07	2.12E+07	2.07E+07
host_id	1.01E+08	5.48E+07	6.62E+07	9.30E+07	9.34E+07
latitude	40.848643	40.685107	40.764935	40.732414	40.610736
longitude	-73.884384	-73.95139	-73.97468	-73.873378	-74.105007
price	86.532443	124.18238	197.52373	98.272011	115.14045
minimum_nights	4.255725	5.932464	8.708857	5.68146	3.926966
number_of_reviews	25.90458	23.906389	20.811978	26.86131	30.446629
reviews_per_month	1.83267	1.279083	1.261122	1.860276	1.799663
calculated_host_listings_count	2.264313	2.401569	12.201147	4.234252	2.328652
availability_365	166.8187	99.441531	111.87843	145.06192	191.76966
neigh	Bronx	Brooklyn	Manhattan	Queens	Staten Island

For table 3, by grouping listings into different room types (Entire home/apt, Private room, and Shared room) and the mean value of each columns, entire home/apt type of room is most expensive with the highest requirement of the average number of minimum nights, private room type of room requires least average number of minimum nights with highest average number of reviews and reviews per month, and lowest average availability over 365 days. Shared room type of room has the lowest average price and the lowest average number of reviews, and the average availability of 365 days is the highest compares to the other two room types.

Table 3: Listing grouped by room type and mean values

room_type	Entire home/apt	Private room	Shared room
id	1.79E+07	1.90E+07	2.24E+07
host_id	5.98E+07	7.03E+07	9.97E+07
latitude	40.729159	40.729108	40.72921
longitude	-73.961123	-73.943258	-73.943106
price	212.8972	89.411517	68.636207
minimum_nights	8.598856	5.40815	6.524138
number_of_reviews	22.701853	23.579552	16.437931
reviews_per_month	1.271822	1.443968	1.514118
calculated_host_listings_count	10.329495	3.267313	5.213793
availability_365	112.19176	110.1	158.56207

2.1.2 *Basic statistics*

From table 4, the minimum price and maximum price are quite irrational. These listings may set extremely high/low price to avoid users to rent their houses because of their reasons. Also, the median shows the price of NYC houses is not very high and 0 is one of the modes shows many listings put an invalid price on Airbnb. To do accurate analysis, analyzed price may need to set in a smaller range, rather than 0 to 10000 USD, but under 800 USD.

Table 4: Basic statistics of the price

Basic statistics of price	Values (USD)
Min	0.00E+00
Max	1.00E+04
Median	107
Mode	0 and 150

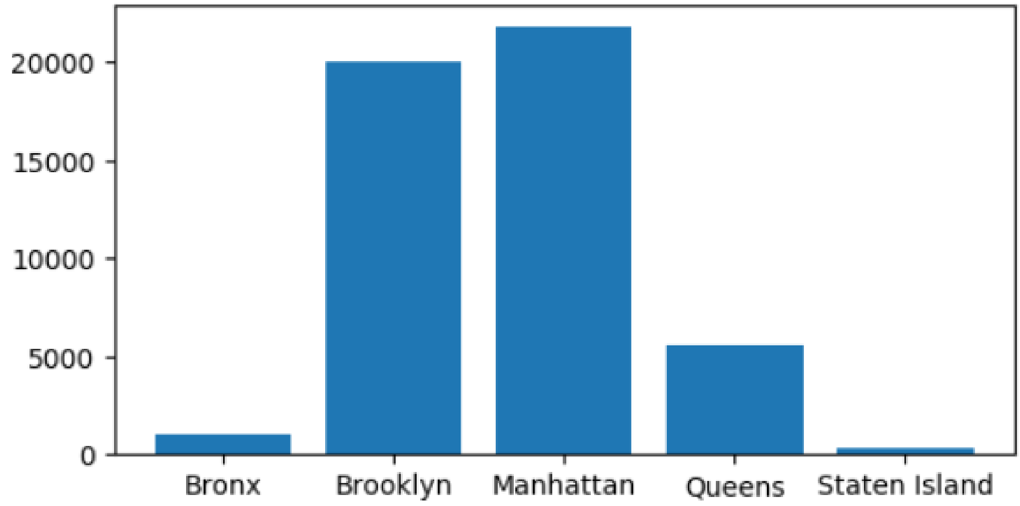

Figure 1: Number of listings in each neighbourhood groups

Figure 1 illustrates the relationship between different neighbourhood groups in data_listing with the number of Airbnb listings in these neighbourhood groups during June 2019. The x-axis is different neighbourhood groups and the y-axis is the number of listings in each neighbourhood groups. From the figure, Manhattan has over 20,000 Airbnb listings which have the top number of houses in NYC. Brooklyn has nearly 20,000 listings, and Staten Island has the lowest number of listings over other neighbourhood groups.

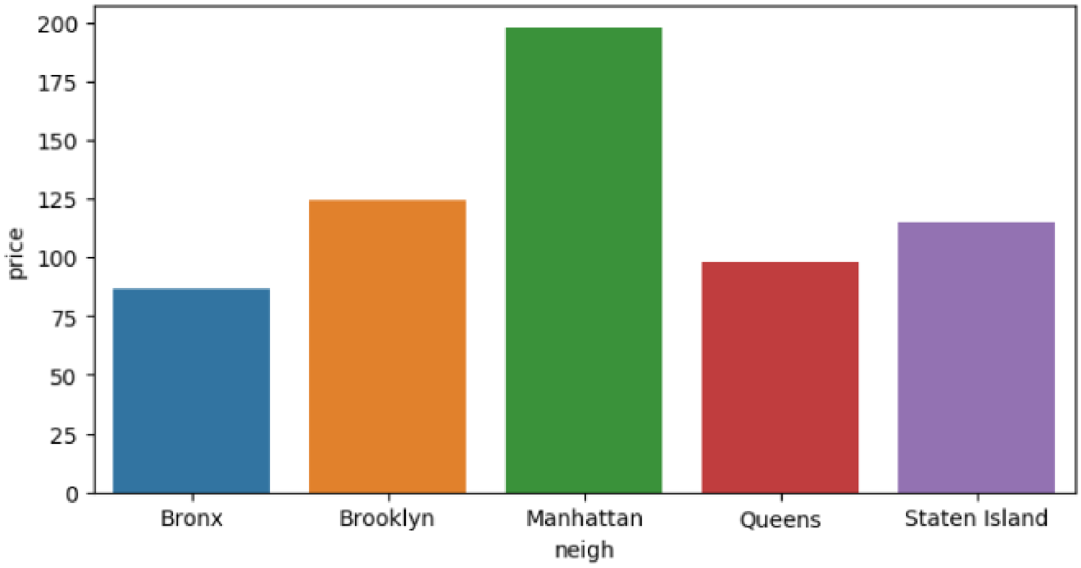

Figure 2: Average price in each neighbourhood groups

Figure 2 shows the relationship between different neighbourhood groups in NYC and average Airbnb house price in these neighbourhood groups during June 2019. The x-axis is different neighbourhood groups and the y-axis is the average price for houses in different neighbourhood groups. From the figure, the Manhattan houses' price is the highest (around 200 USD per day). Brooklyn, Staten Island, and Queens follows up. Bronx houses' price is the lowest compares with other neighbourhood groups. In addition, the reality tells that Manhattan houses are more expensive than other areas in NYC, so the relationship on the figure is quite accurate.

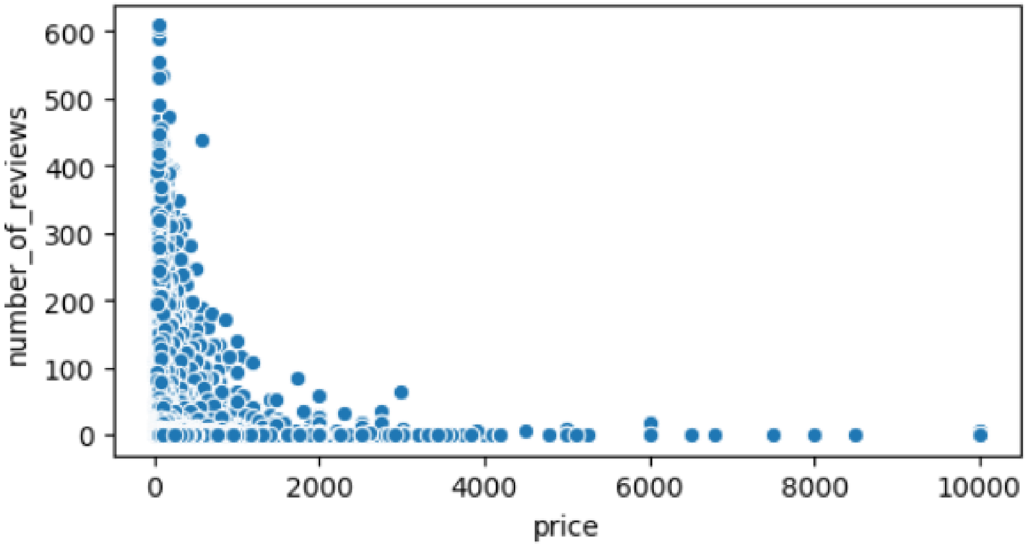

Figure 3: Number of reviews in different price range

Figure 3 demonstrates the relationship between each Airbnb houses' price in NYC during June 2019 with the number of reviews for each listing. The x-axis is the house price and the y-axis is the number of reviews for each listing that have a different price range. From the figure, the number of review and house price are in an inverse proportion that listings with lower houses' price have more number of reviews. In addition, Airbnb houses' price in NYC during June 2019 are mostly under 1000 USD. For the accuracy of data analysis, it is reasonable to eliminate the range of "price" to 800 USD in order to remove the effect of outliers (price>800).

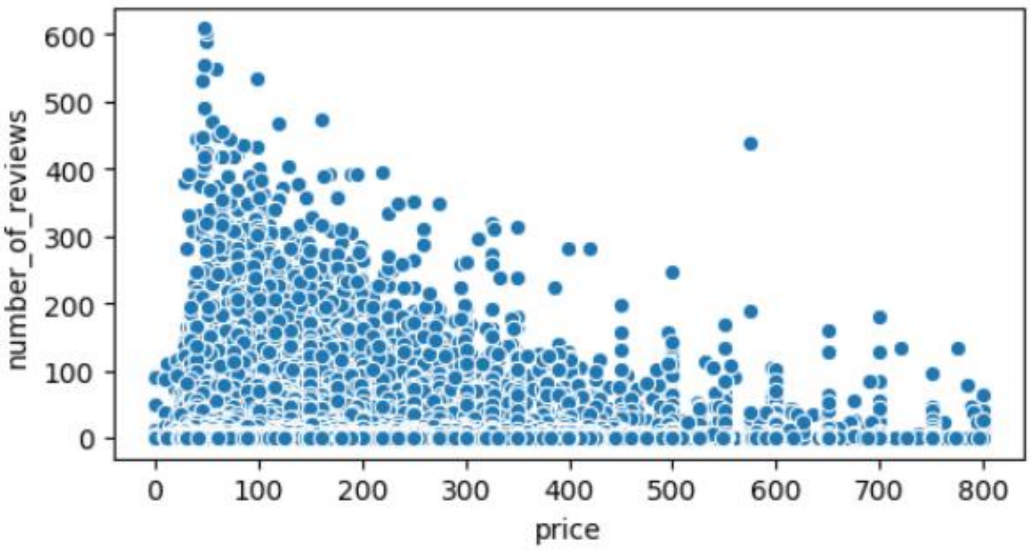

Figure 4: Relationship between number of reviews and price

Figure 4 shows the relationship between each Airbnb houses' price (under 800 USD) with the number of reviews for each listing in NYC during June 2019. The x-axis is the houses' price and the y-axis is the number of reviews for each listing for different houses' price. By eliminating the range of price, it clearly illustrates an inverse proportionally relationship between the number of reviews and houses' price. Moreover, figure 4 shows the number of Airbnb listings with the price around 0 USD to 300 USD is quite big, so many of Airbnb listings in this data set are around this price range.

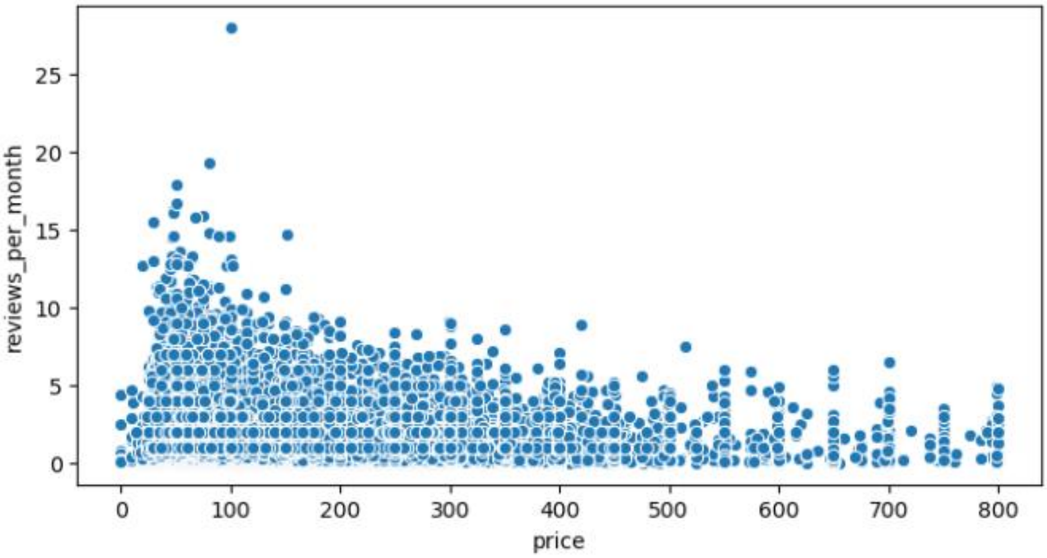

Figure 5: Relationship between price and reviews per month

Figure 5 shows the relationship between Airbnb house price (under 800 USD) in NYC during June 2019 and reviews per month (June 2019) for each listing. The x-axis is the house price and the y-axis is numbers of reviews per month for each listing with different houses' price. There is a slightly inverse proportional relationship between price and reviews per month. There are also a few outliers illustrated in the figure, but the overall relationship does not influence much by that.

2.2 Listing detail dataset

2.2.1 Data overview

Listing detail data frame (data_listing_detail) is the detailed information of Airbnb listings in New York City during June 2019. There are 48801 rows and 106 columns in this data frame. Information including listing id, the summary about the house, space description about the house, cancellation policy, guest requirements, reviews about the house, hosts' information, etc. It offers more detailed information than the data_listing data frame.

2.2.2 Basic statistics

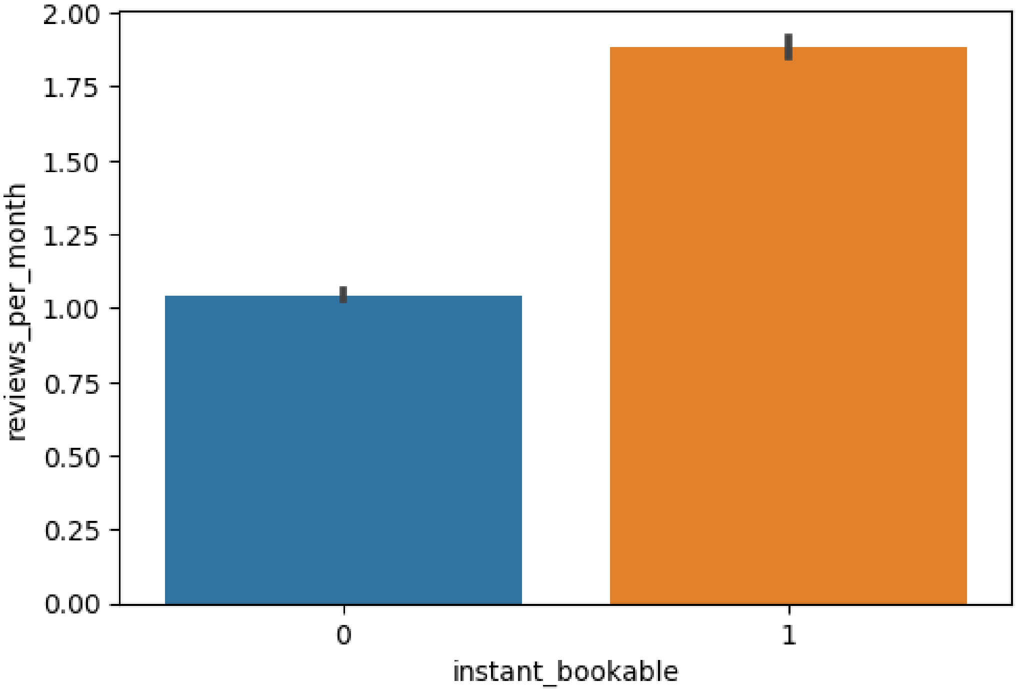

Figure 6: Relationship between instant bookable and reviews per month

Figure 6 shows the relationship between the number of Airbnb houses in NYC during June 2019 that are both instant bookable or not with the number of review per month. The x-axis shows whether the listing is instantly bookable or not (0: not instant bookable, 1: instant bookable) and the y-axis shows reviews per month for listings that are both instant bookable and not instant bookable. From the figure, it demonstrates instantly bookable listings have more reviews than those are not instant bookable. Therefore, users tend to choose listings that are easier to access and can book immediately.

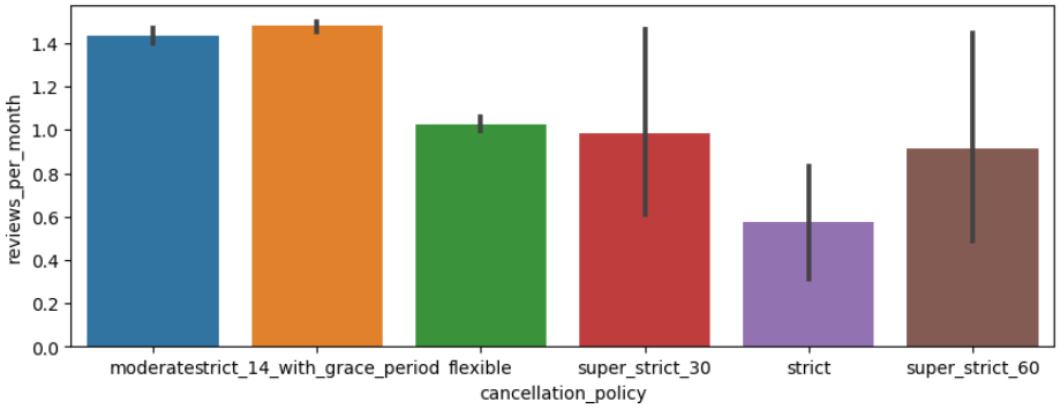

Figure 7: Relationship between cancellation policy and reviews per month

Figure 7 shows the relationship between different cancellation policies (flexible: full refund within 1 day prior, moderate: full refund within 5 day prior, strict_14_with grace period: full refund if cancelled in the first two days of booking as long as the booking is more than 14 days in future, super_strict_30: 50% refund up until 30 days prior to check-in, minus service fees, strict: full refund if cancellation is within 48 hours of booking, super_strict_60: 50% refund up until 60 days prior to check-in, minus service fees) and reviews per month for Airbnb listings in NYC during June 2019. The x-axis is different cancellation policy and the y-axis is the number of reviews per month for these listings with the different cancellation policy. From the figure, strict 14 with grace period has the most reviews per month, moderate follows up, and strict has the least reviews per month. It can infer that people are more prefer to book a listing with strict 14 with a grace period and moderate cancellation policy because they can easily change their booking plan whenever they want without loses. It is quite surprising that super strict 60 has more reviews per month than strict, but these reviews may contain some negative expressions and that is reasonable.

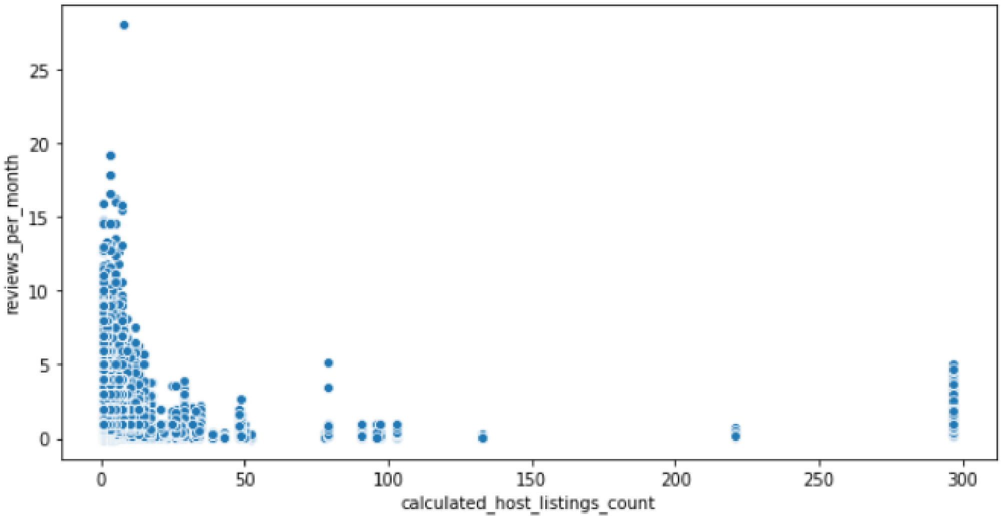

Figure 8: Relationship between calculated host listings count and reviews per month

Figure 8 shows the relationship between calculated host listings count for Airbnb listings in NYC during June 2019 and reviews per month for these listings. The x-axis is the calculated host listings count and the y-axis is the number of reviews per month for listings. The relationship between these two variables is not very clear, but it shows a slightly inverse proportional relationship. From the figure, the host that has fewer listings has more reviews per month, however, there are still some outliers. Besides, hosts may put invalid information onto the website, for example, hosts with 250 and more listings probably do not have that many listings or most of the listings are not available for users. Therefore, the relationship can be largely change by this.

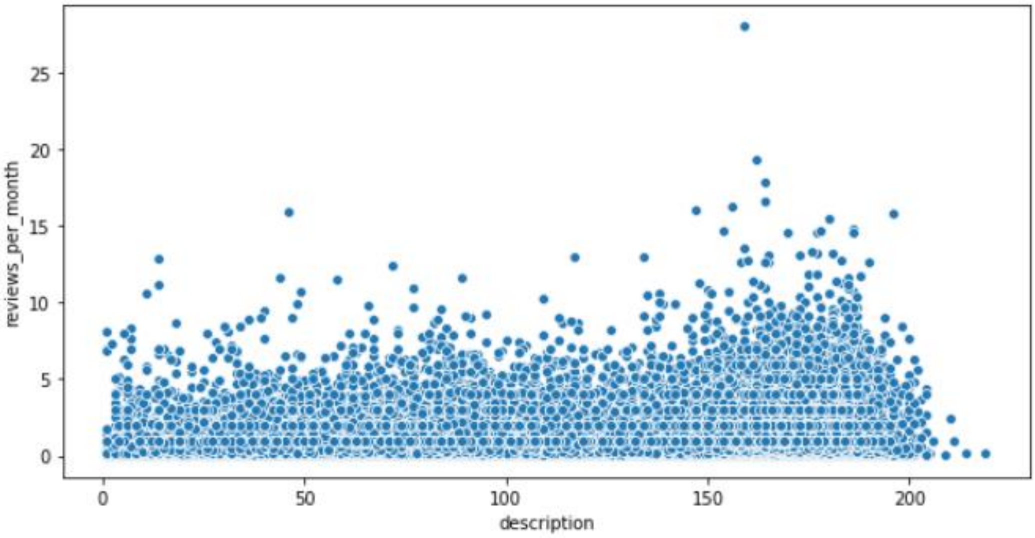

Figure 9: Relationship between length of description and reviews per month

Figure 9 shows the relationship between the length of description for Airbnb listings in NYC during June 2019 and reviews per month for these listings. The x-axis is the length of description and the y-axis is reviews per month. From the figure, it illustrates a directly proportional relationship that the longer the description the listing has, the more reviews it gets per month, but the relationship is not very obvious. Moreover, since NYC is a diverse city, hosts may come from different countries and the language in the description may be different. Since different languages express meanings in different ways, the length may be influenced by that.

2.3 Calendar dataset

2.3.1 Data overview

Calendar data frame (data_calendar) contains information of Airbnb listings in New York City from June 2019 to June 2020. There are 17812547 rows and 7 columns in this data frame. Total rows are calculated using 48801 (total number of listings) times 365/366 days. Information including listing id, dates, price, adjusted price, availability during specific dates, minimum nights and maximum nights.

2.3.2 Basic statistics

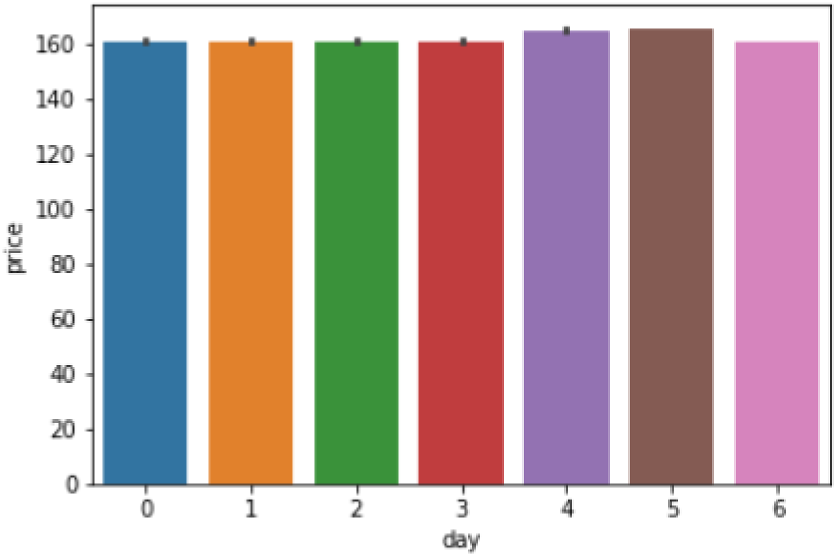

Figure 10: Relationship between day of week and the house price in next 365 days

Figure 10 illustrates the relationship between day of the week and the Airbnb house price in NYC for the next 365 days. From the figure, during Friday (4) and Saturday (5), the price is higher than other days. That is probably because of people are having a vacation using Friday and Saturday, because they are working on other weekdays and get back to work using Sunday, so other days' price is lower than Friday and Saturday.

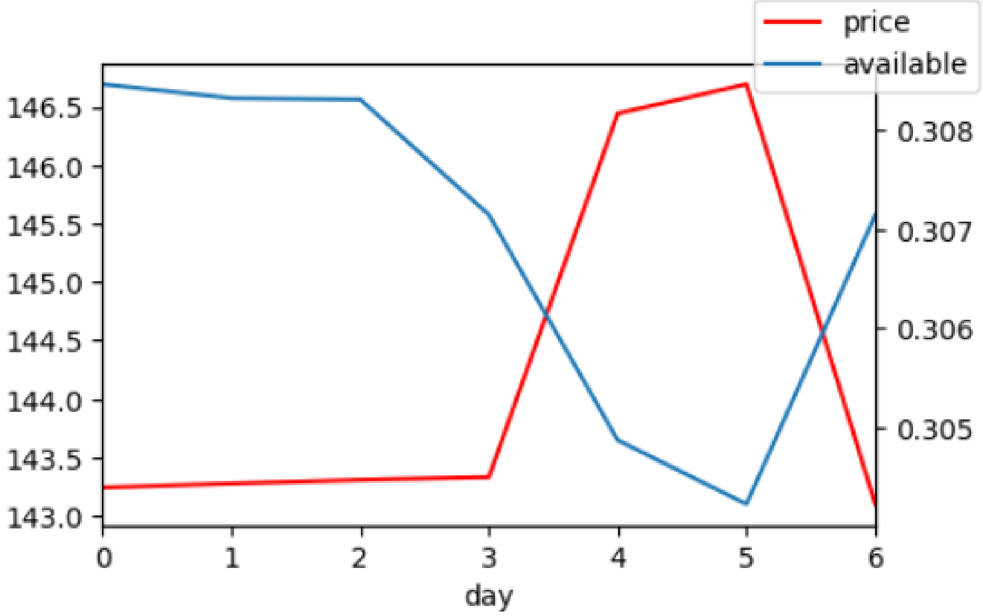

Figure 11: Average price and availability in each day

Figure 11 illustrates the relationship between average NYC Airbnb house price and availability for the next 365 days with days. The figure can demonstrate the relationship that during Friday and Saturday, the house price is higher. And during Friday and Saturday, the availability is low maybe because of people's demand. For Sunday, the price is lower and the availability gets back to a higher value. For weekdays, the price is very moderate and the availability is stable from Monday to Wednesday, drops during Thursday.

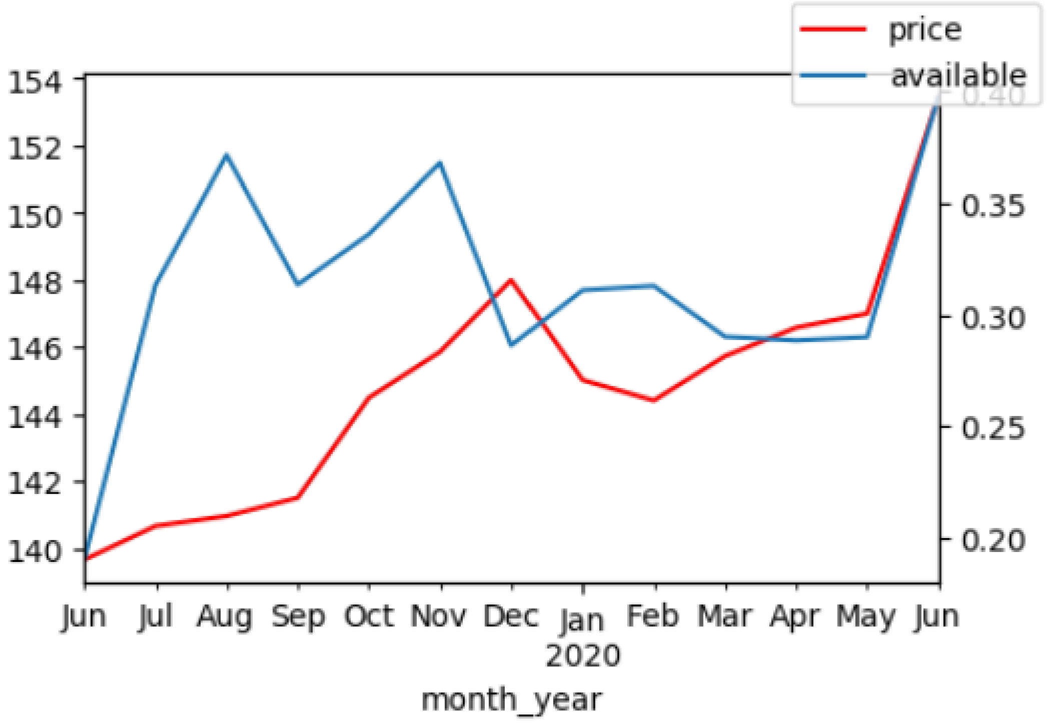

Figure 12: Average price and availability from June 2019 to June 2020

Figure 12 illustrates the relationship between average NYC houses' price and availability from June 2019 to June 2020. From the figure, the overall trend of the house price is growing. During December 2019, the price is the highest in 2019 maybe because of Christmas and other holidays. For June 2020, the house price is also quite high probably because of the tourist season. For the availability, the slope is very unstable, with high availability during July 2019, November 2019 and June 2020.

Table 5: Basic statistics for maximum night and minimum night

Basic Statistics	Values
Minimum Nights Min	1.00E+00
Minimum Night Max	5.04E+03
Maximum Night Min	1
Maximum Night Max	2147483647
Minimum Night Median	3
Maximum Night Median	804
Minimum Night Mean	11.17
Maximum Night Mean	45970.15

Table 5 shows the maximum value, minimum value, median, and mean of minimum night and maximum night for each Airbnb houses in NYC from June 2019 to June 2020. From the table, the maximum value of the minimum night and the maximum value of the maximum night are unrealistic. Therefore, they may lead the result to become inaccurate.

2.4 Data visualization

2.4.1 Time series of price history

Grouped by neighbourhood group

Figure 13 shows the relationship between the average price of Airbnb listings in NYC in 5 neighbourhood groups and time. From the figure, Manhattan houses have the greatest houses' price among other neighbourhood groups from January 2015 to July 2019. Bronx houses' price is the lowest. The houses' price in Staten Island is very unstable, the highest point is 185 and the lowest point is 110, and the overall relationship of houses' price on Staten island with time is decreasing.

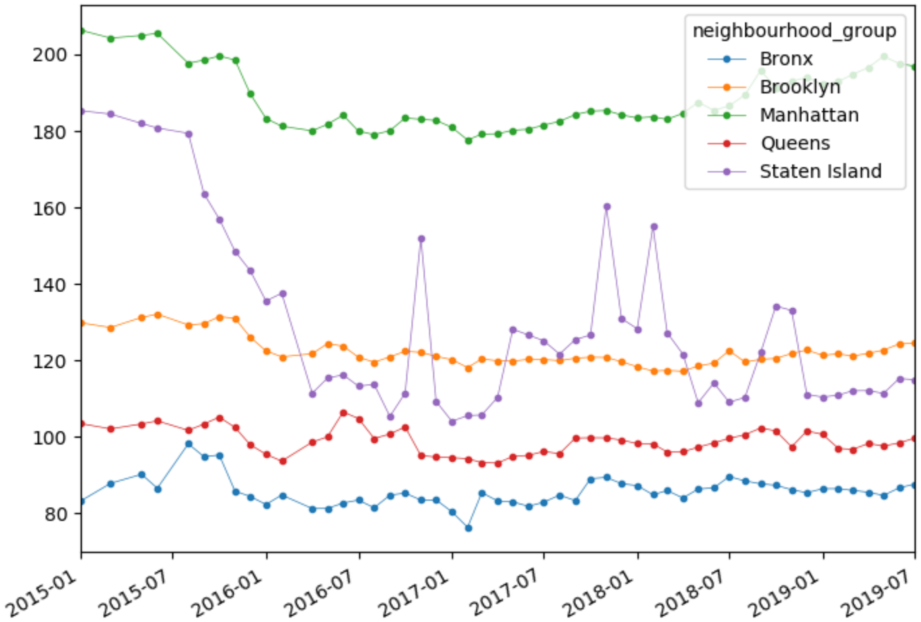

Figure 13: Average price from Jan 2015 to July 2019

Figure 14 shows the relationship between average houses' count of Airbnb listings in NYC in 5 neighbourhood groups and time. From the figure, Manhattan houses have the greatest houses' count among other neighbourhood groups from January 2015 to July 2019. Staten Island has the lowest amount of houses, following by Bronx, Queens, and Brooklyn. The number of houses in five neighbourhood groups are growing through these years but Manhattan houses count has a decreasing trend from the beginning of 2019, as well as Brooklyn.

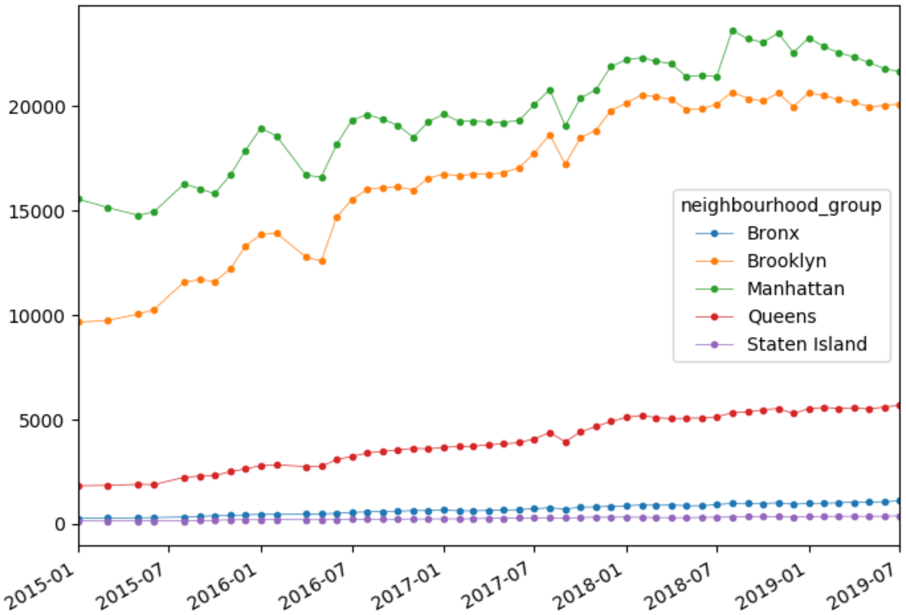

Figure 14: Count from Jan 2015 to July 2019

Figure 15 shows the relationship between average houses' availability in 365 days of Airbnb houses in NYC in 5 neighbourhood groups and time. From the figure, Staten Island has the highest availability among other neighbourhood groups' listings, following by Bronx, Queens, Brooklyn, and Manhattan. During late 2018, Manhattan houses' availability was growing and it was higher than

Brooklyn houses' availability. The overall trend of houses' availability in these five neighbourhood groups is decreasing.

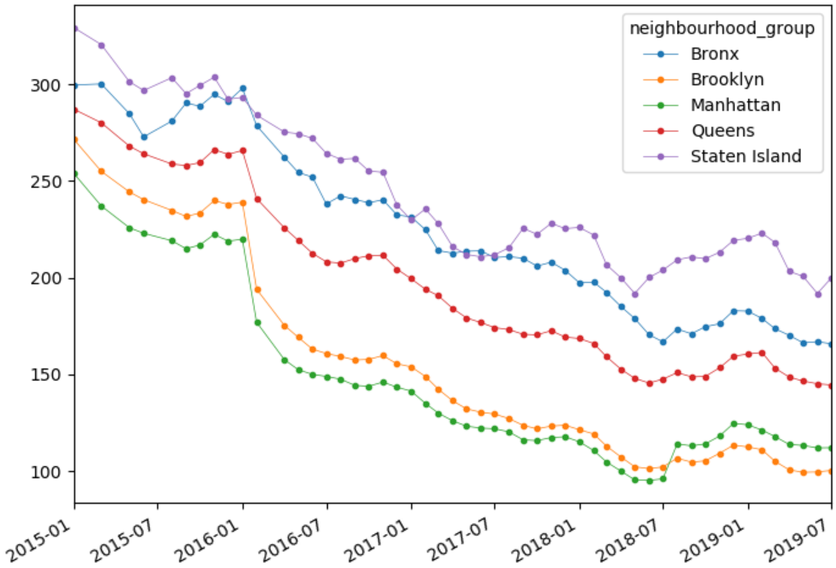

Figure 15: Average availability in 365 days from Jan 2015 to July 2019

Grouped by room type

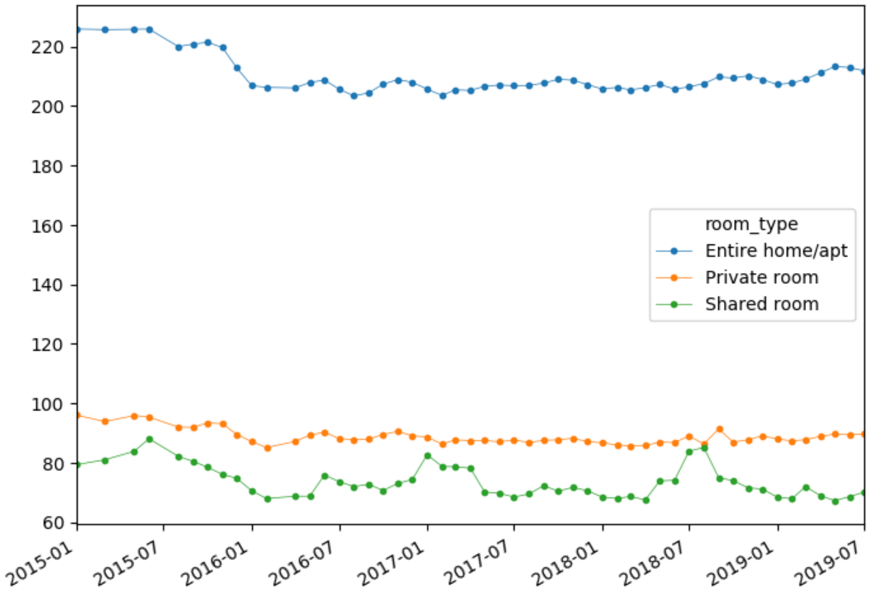

Figure 16: Count from Jan 2015 to July 2019

Figure 16 shows the relationship between average houses' count of Airbnb listings in NYC with different room types and time. From the figure, the count of the entire home/apt and private room are growing with time. For the count of the private room, the trend is very gentle (a gently growing trend) and the count is always around 1000 from 2015 to 2019.

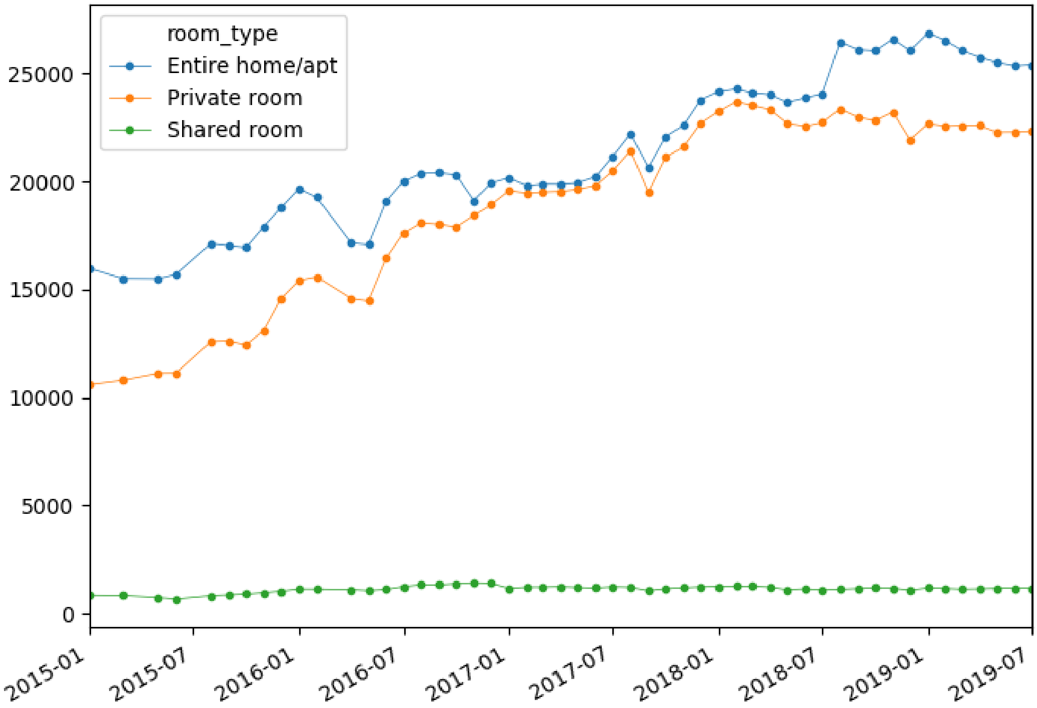

Figure 17: Count from Jan 2015 to July 2019

Graph 17 shows the relationship between average houses' count of Airbnb listings in NYC with different room types and time. From the graph, the count of the entire home/apt and private room are growing with time. For the count of the private room, the trend is very gentle (a gently growing trend) and the count is always around 1000 from 2015 to 2019.

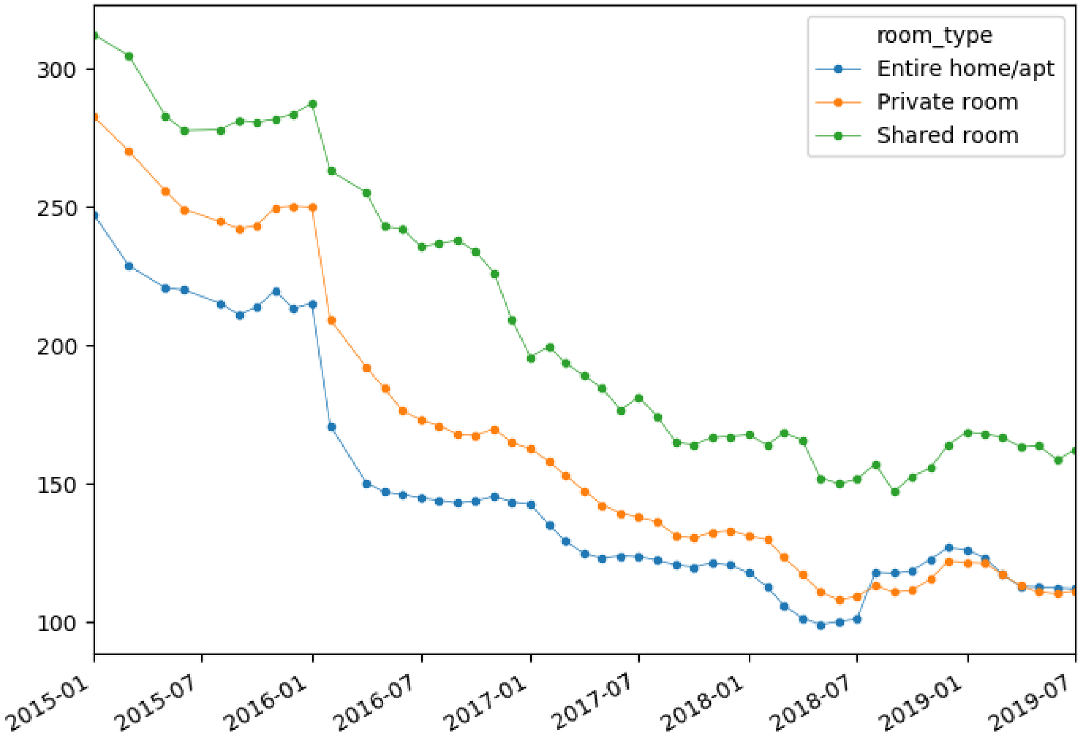

Figure 18: Average availability in 365 days from Jan 2015 to July 2019

Figure 18 shows the relationship between average houses' availability in 365 days of Airbnb listings in NYC for different room types and time. From the figure, the shared room has more availability than the private room and the entire home/apt. The overall trend of entire home/apt, private room and the shared room are decreasing, the availability of houses for these room types decreases while the time goes on.

2.4.2 Map of locations with features

Map of locations with house density

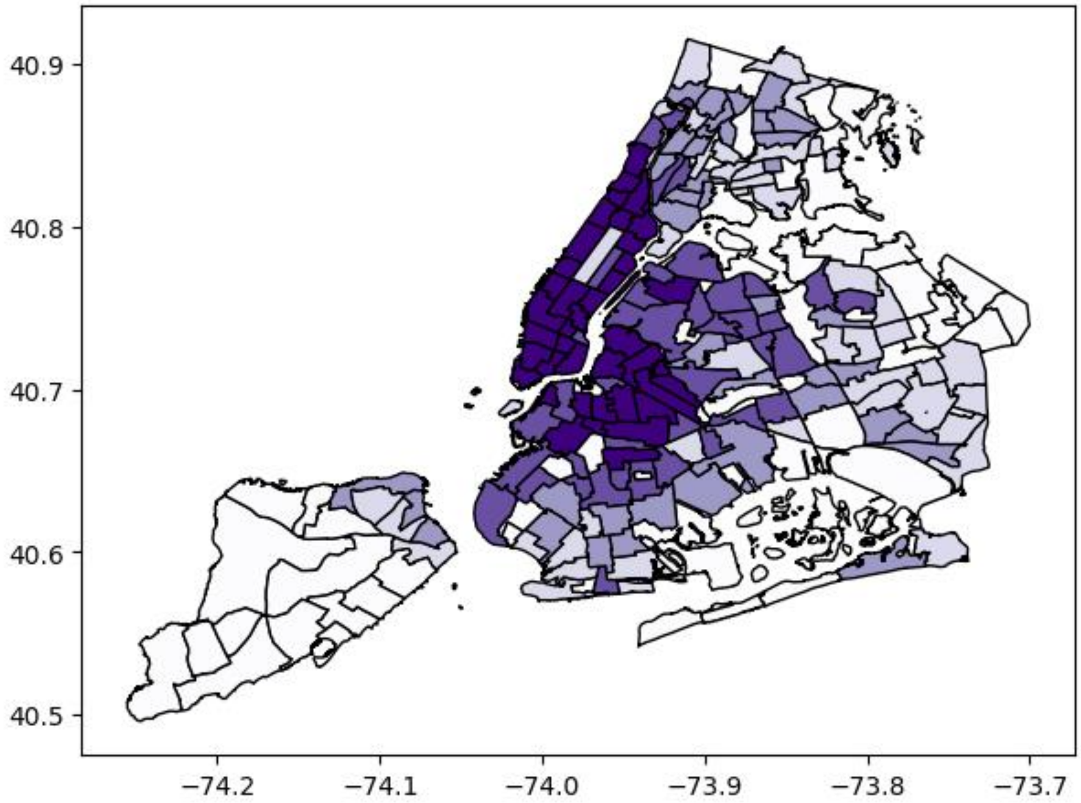

Figure 19: Listing density in NYC in June 2019

Figure 19 shows the relationship between house density of NYC during June 2019 with different areas in NYC. The darker the purple is, the higher the listing density in this area is. From the figure, Manhattan's house density is very high, shown by the dark purple. And Brooklyn house density is high as well. However, for Staten Island, Bronx and Queens have lower house density, shown by the light purple color. Since the tourism in the Manhattan area is better than other areas and the population is larger, it is reasonable that the listing density is higher.

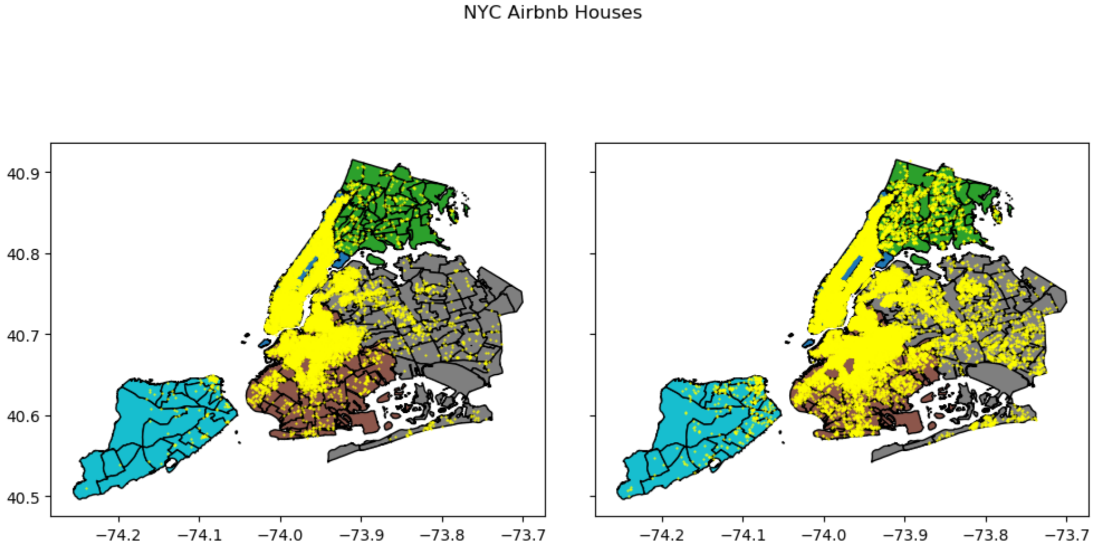

Figure 20: Airbnb listings in NYC in Jan 2015 and in June 2019

Figure 20 shows the difference between the number of houses in NYC in January 2015 and in June 2019. The figure on the left is the house count in NYC during January 2015, the figure on the right is the house count in NYC during June 2019. There is a great difference between the house count that during June 2019 there are more houses than January 2015. Especially in Bronx, Brooklyn and in Queens, the house count grows significantly.

Map of locations with availability

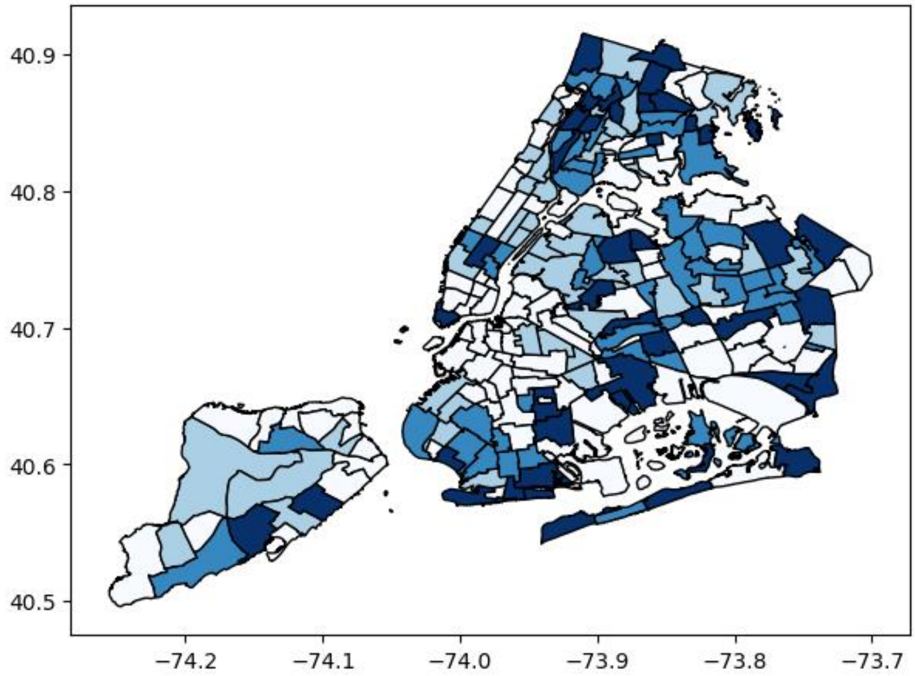

Figure 21: House availability in NYC in June 2019

Figure 21 shows the relationship between house availability of NYC during June 2019 with different areas in NYC. The darker the blue is, the higher the house availability in this neighbourhood tabulation area is. From the figure, Manhattan's house availability is lower than in other areas, shown by the light blue color. Brooklyn house availability is high, shown by the dark blue color. For Bronx and Queens, the house availability is moderate, but for Staten Island, the house availability is low.

Map of locations with price

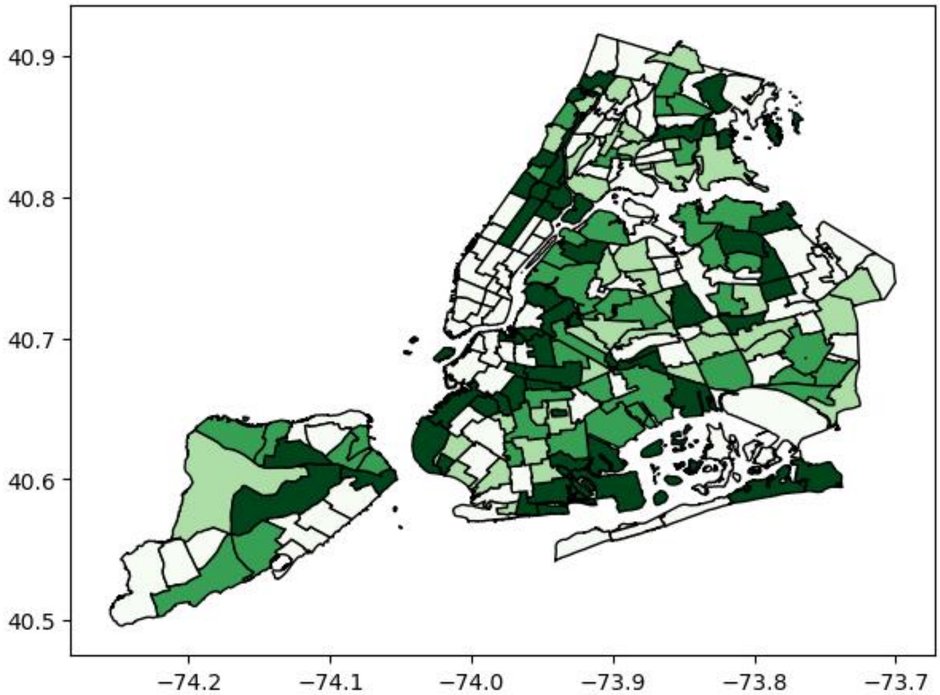

Figure 22: House price in NYC in June 2019

Figure 22 shows the relationship between house price of NYC during June 2019 with different areas in NYC. The darker the green is, the higher the house price in this neighbourhood tabulation area is. From the figure, Manhattan's houses' price is significantly higher than other neighbourhood tabulation area, the houses' price in Bronx and Brooklyn is quite moderate, the houses' price in Queens and Staten Island is low, shown by the lightness of the color.

3. Model

3.1 Regression analysis

Table 6 is the linear regression result based on numbers of review, availability in 365 days, reviews per month, minimum nights, length of description, requirement of guest phone verification, cancellation policy, neighbourhood group, host response rate, cleaning fee, accommodates, room types of houses, and property type of listings, but eliminates the P values bigger than 5%. The R squared value is 0.559, the dependent variable is price, the number of observation is 24104. For more accurate analysis, variables that have P-value bigger than 5% should be removed.

Table 6: Linear regression result for Airbnb houses' price in June 2019

Dep. Variable:	price	R-squared:	0.559
Model:	OLS	Adj. R-squared:	5.58E-01
Method:	Least Squares	F-statistic:	6.94E+02
Date:	Tue, 03 Sep 2019	Prob (F-statistic):	0
Time:	14:57:56	Log-Likelihood:	-1.37E+05
No. Observations:	24104	AIC:	2.74E+05
Df Residuals:	24059	BIC:	2.74E+05
Df Model:	44		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.8402	1.859	6.907	0	9.197	16.484
number_of_reviews	-0.0902	0.01	-9.125	0	-0.11	-0.071
availability_365	0.0382	0.003	11.516	0	0.032	0.045
reviews_per_month	-1.3812	0.296	-4.663	0	-1.962	-0.801
minimum_nights	-0.375	0.025	-15.298	0	-0.423	-0.327
super_strict_60	44.3834	14.855	2.988	0.003	15.267	73.5
Brooklyn	12.2886	1.342	9.158	0	9.659	14.919
Manhattan	60.2693	1.402	42.996	0	57.522	63.017
Staten Island	-26.2657	4.55	-5.773	0	-35.184	-17.347
Bronx	-12.2696	2.887	-4.25	0	-17.928	-6.611
cleaning_fee	0.4017	0.01	39.883	0	0.382	0.421
accommodates	11.8809	0.324	36.664	0	11.246	12.516
Entire home/apt	50.5932	1.011	50.046	0	48.612	52.575
Shared room	-20.8758	2.754	-7.58	0	-26.274	-15.478
bedrooms	24.1465	0.736	32.823	0	22.705	25.588
Resort	370.9991	12.584	29.482	0	346.334	395.664
Boutique hotel	83.7515	10.355	8.088	0	63.455	104.048

House	-18.8458	1.811	-10.404	0	-22.396	-15.295
Guest suite	-31.8766	4.465	-7.14	0	-40.627	-23.126
Apartment	-23.3958	1.228	-19.059	0	-25.802	-20.99
Hostel	-19.7385	9.787	-2.017	0.044	-38.921	-0.556

Table 7 is the linear regression model for Airbnb houses' price in NYC during June 2019 with independent variables have P-value bigger than 5% removed. For this linear regression, the dependent variable is the house price, and the independent variables are number of reviews, availability of house in 365 days, reviews per month, minimum nights, super strict 60 cancellation policy, Brooklyn, Manhattan, Staten Island, Bronx, cleaning fee, numbers of accommodation for the house, room types (entire home/apt and shared room), and proper types (bedrooms, resort, boutique hotel, house, guest suite, apartment, hostel). The number of observation is 31983 and the R-squared is 0.541, which shows the range of the big data is large.

Table 7: Linear regression of Airbnb houses' price in June 2019 (P value>5% removed)

Dep. Variable:	price	R-squared:	0.541
Model:	OLS	Adj. R-squared:	0.541
Method:	Least Squares	F-statistic:	1883
Date:	Mon, 19 Aug 2019	Prob (F-statistic):	0
Time:	10:46:03	Log-Likelihood:	-1.81E+05
No. Observations:	31983	AIC:	3.62E+05
Df Residuals:	31962	BIC:	3.62E+05
Df Model:	20		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	12.8402	1.859	6.907	0	9.197	16.484
number_of_reviews	-0.0902	0.01	-9.125	0	-0.11	-0.071
availability_365	0.0382	0.003	11.516	0	0.032	0.045
reviews_per_month	-1.3812	0.296	-4.663	0	-1.962	-0.801
minimum_nights	-0.375	0.025	-15.298	0	-0.423	-0.327
super_strict_60	44.3834	14.855	2.988	0.003	15.267	73.5
Brooklyn	12.2886	1.342	9.158	0	9.659	14.919
Manhattan	60.2693	1.402	42.996	0	57.522	63.017
Staten Island	-26.2657	4.55	-5.773	0	-35.184	-17.347
Bronx	-12.2696	2.887	-4.25	0	-17.928	-6.611
cleaning_fee	0.4017	0.01	39.883	0	0.382	0.421
accommodates	11.8809	0.324	36.664	0	11.246	12.516
Entire home/apt	50.5932	1.011	50.046	0	48.612	52.575
Shared room	-20.8758	2.754	-7.58	0	-26.274	-15.478
bedrooms	24.1465	0.736	32.823	0	22.705	25.588
Resort	370.9991	12.584	29.482	0	346.334	395.664
Boutique hotel	83.7515	10.355	8.088	0	63.455	104.048
House	-18.8458	1.811	-10.404	0	-22.396	-15.295
Guest suite	-31.8766	4.465	-7.14	0	-40.627	-23.126
Apartment	-23.3958	1.228	-19.059	0	-25.802	-20.99
Hostel	-19.7385	9.787	-2.017	0.044	-38.921	-0.556

For the number of reviews, the coefficient is -0.0902, which shows the higher the number of reviews is, the lower the house's price is. The relationship is inversely proportional.

For availability in 365 days, the more available the house is, the higher the house's price is, but the relationship is very slight, shown by the small coefficient 0.0382.

For reviews per month, the higher reviews per month, the lower the houses' price since the coefficient is -1.3812. For the minimum nights, the higher requirement of the number of minimum nights, the lower the house price, shown by the coefficient -0.3750.

For the cancellation policy super strict 60, if the house has the cancellation policy of super strict 60, the house price is higher than others, and there is a directly proportional relationship.

For houses in Brooklyn neighbourhood group, the houses' price is higher, shown by the coefficient 12.2886. For houses in Manhattan, the houses' price is much higher than in other areas, shown by the coefficient 60.2693. For houses in the Bronx, the houses' price is lower indicated by the coefficient -12.2696. For houses in Staten Island, the houses' price is much lower than other area's houses because of the coefficient -26.2657.

For the cleaning fee, the higher the cleaning fee, the higher the houses' price shown by the coefficient 0.4017.

For the accommodates number for the house, the higher of the number of accommodates for the house, the higher the house price, shown by the coefficient 11.8809.

For the room types, if the house's room type is shared a room, the house price is lower, if the house's room type is entire home/apt, the house price is higher (shown by the coefficient 50.5932 and -20.8758).

For the property types, if the house is a bedroom, the house price is higher because of the coefficient 24.1465. If the house is a resort, the price is much higher because the coefficient is higher than the other room types'. If the house's room type is the boutique hotel, the house price is higher than the bedroom but not resort because of the coefficient 83.7515. If the house's room type is a house, the house price is lower since the coefficient is -18.8458. If the house's room type is a guest suite, the house price is lower since the coefficient is -31.8766 and the coefficient is lower than the house's (room type) coefficient. And if the room type of the house is an apartment, the price is lower, as well as a hostel, shown by the coefficient -23.3958 and -19.7385.

3.2 Time series analysis

Table 8 is the linear regression model for Airbnb houses' price of the calendar from June 2019 to June 2020. The dependent variable is the price for the calendar data frame, the independent variables are minimum nights for listing data frame, number of accommodates for each house, cleaning fee, availability in 365 days, number of reviews, instant bookable, review per month, room types, neighbourhood groups, days of week and month with years. The number of observation is 116446 because the data for the calendar for listings is too big. A sample of 1% of the whole data_calendar was used for sample observation. The R-squared for the linear regression is 0.521.

For independent variables that are not related to time and calendar, the explanations were shown above. For independent variables that are related to time and calendar, if the day is Monday, the house price is lower, shown by the coefficient -1.1817. If the day is Tuesday, the price has a slightly direct proportional trend (shown by the coefficient 0.049). If the day is Wednesday, the price also has a slightly direct proportional trend (shown by the coefficient 0.0791). If the day is Thursday, the price has an inverse proportional slope (shown by the coefficient -0.1088). If the day is Friday, the price has a strong direct proportional trend because of the big coefficient 3.5807. If the day is Saturday, the price has a greater direct proportional trend because of the coefficient 4.5526. For Sunday, the house price has a slightly inverse proportional trend (shown by the coefficient -0.8137).

For June 2019, July 2019, August 2019, September 2019 and October 2019, the relationship between the price is inversely proportional. For November 2019, December 2019, January 2020, February 2020, March 2020, April 2020, May 2020 and June 2020, the relationship between the price is directly proportional. Notably for November, December of 2019, and March to June of 2020, the coefficient is bigger than other months, indicating the seasonality of higher price during November and December because of holidays and higher price during March to June.

Table 8: Linear regression of Airbnb houses' price of calendar during June 2019 to June 2020

Dep. Variable:	**price_cal**	**R-squared:**	**0.521**
Model:	OLS	Adj. R-squared:	0.521
Method:	Least Squares	F-statistic:	4079
Date:	Tue, 03 Sep 2019	Prob (F-statistic):	0
Time:	21:38:31	Log-Likelihood:	-6.65E+05
No. Observations:	116446	AIC:	1.33E+06
Df Residuals:	116414	BIC:	1.33E+06
Df Model:	31		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	6.813	0.545	12.512	0	5.746	7.88
minimum_nights_list	-0.3473	0.012	-28.1	0	-0.372	-0.323
accommodates	20.461	0.141	145.359	0	20.185	20.737
cleaning_fee	11:43:26	0.005	8.93E+01	0	0.478	0.499
availability_365	0.0507	0.002	2.80E+01	0	0.047	0.054
number_of_reviews	-0.0923	0.005	-1.69E+01	0	-0.103	-0.082
instant_bookable	-0.845	0.462	-1.829	0.067	-1.75	0.06
reviews_per_month	-0.4451	0.168	-2.653	0.008	-0.774	-0.116
Private room	-4.2878	0.491	-8.729	0	-5.251	-3.325
Entire home/apt	38.1032	0.561	67.893	0	37.003	39.203
Shared room	-27.0024	1.108	-24.375	0	-29.174	-24.831
Brooklyn	10.0692	0.586	17.188	0	8.921	11.217
Manhattan	55.2582	0.597	92.636	0	54.089	56.427
Queens	-5.5452	0.728	-7.616	0	-6.972	-4.118
Staten Island	-32.6725	1.993	-16.391	0	-36.579	-28.766
Bronx	-20.2967	1.278	-15.88	0	-22.802	-17.792
3	-0.1088	0.53	-0.205	0.837	-1.148	0.93
1	0.7049	0.531	1.327	0.185	-0.337	1.746
6	-0.8137	0.529	-1.537	0.124	-1.851	0.224
4	3.5807	0.533	6.72	0	2.536	4.625
0	-1.1817	0.53	-2.229	0.026	-2.221	-0.143
5	4.5526	0.534	8.519	0	3.505	5.6
2	0.0791	0.532	0.149	0.882	-0.964	1.122
19-Dec	3.7682	0.777	4.851	0	2.246	5.291
19-Oct	-0.3117	0.779	-0.4	0.689	-1.839	1.216
19-Nov	2.5236	0.792	3.188	0.001	0.972	4.075
19-Sep	-3.0554	0.789	-3.872	0	-4.602	-1.509
19-Jul	-1.7548	0.777	-2.258	0.024	-3.278	-0.232
20-May	3.7943	0.782	4.855	0	2.262	5.326
19-Aug	-1.6966	0.777	-2.183	0.029	-3.22	-0.173
20-Jan	0.5355	0.789	0.679	0.497	-1.01	2.081
20-Apr	2.0641	0.792	2.608	0.009	0.513	3.615

19-Jun	-4.7052	0.809	-5.818	0	-6.29	-3.12
20-Mar	2.1063	0.78	2.702	0.007	0.578	3.634
20-Feb	0.7023	0.809	0.868	0.385	-0.883	2.288
20-Jun	2.8424	4.209	0.675	0.499	-5.407	11.092

Table 9 shows the final prediction of the coefficient for different days and months with years. From the table, Friday and Saturday's coefficient is bigger than other days and November 2019, December 2019, March 2020 to June 2020 has bigger coefficient as well, indicating the seasonality mentioned before.

Table 9: Final coefficient for day and month_year

Independent Variables	**Coefficient**
Monday	-1.3104788
Tuesday	0.6480953
Wednesday	-0.2003785
Thursday	-0.294686
Friday	3.4471084
Saturday	4.4411362
Sunday	-0.7324911
19-Jun	-4.9474829
19-Jul	-1.9292282
19-Aug	-1.6500932
19-Sep	-3.0976451
19-Oct	-0.5000482
19-Nov	2.1852175
19-Dec	3.8856216
20-Jan	0.7934693
20-Feb	0.4836114
20-Mar	2.3232776
20-Apr	1.7872608
20-May	3.842154
20-Jun	2.8221908

Figure 23 indicates the directly proportional relationship between days of weeks and the value of the coefficient. During Friday and Saturday, the value is high and the slope is rising to a large extent because of the weekend. During Sunday, the coefficient value drops because people need to go back to work for weekdays.

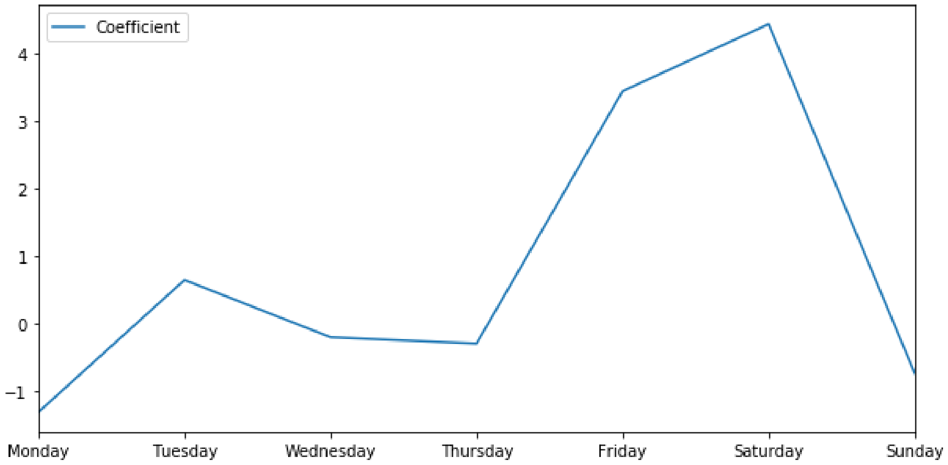

Figure 23: Relationship between days and the value of the coefficient

In figure 24, the slope is positive by months, especially during October_2019 to December 2019. The reason is that most of the American holidays are in this period, so the amount of tourists grows and citizens come back home with their families. During the springtime of 2020, the value of coefficient keeps rising but finally end up with a decreasing trend in June 2020.

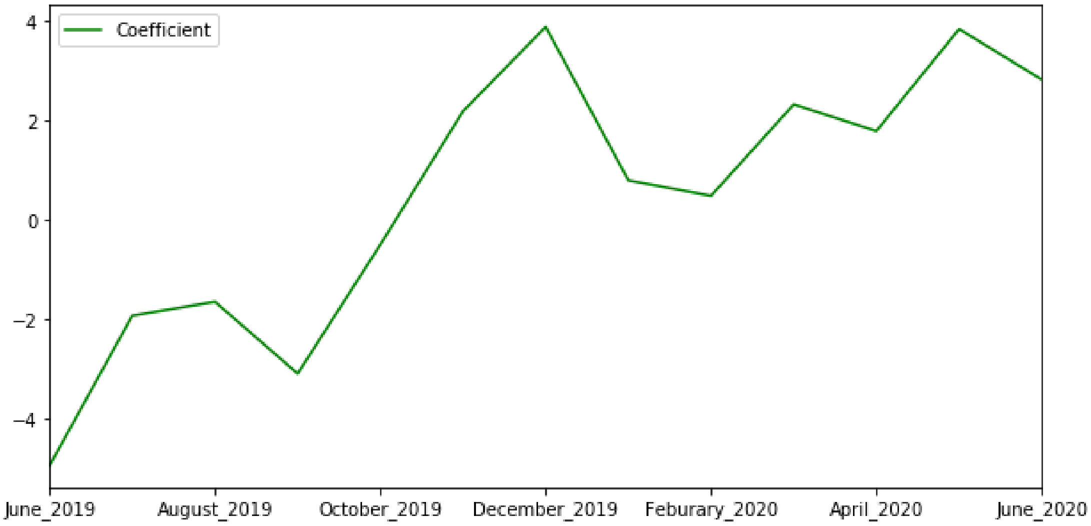

Figure 24: Relationship between months with tears and the value of the coefficient

3.3 Prediction

Table 10 illustrates the mean squared error for the training set and the mean squared error for the testing set for the linear regression of NYC Airbnb houses' price during June 2019 for different test size (from 0.1 to 0.5). From the table, there is no big difference between the train mean squared error and test mean squared error, indicating the prediction is quite accurate. And for test size 0.1, the difference between train mean squared errors and test mean squared error is bigger than other test sizes because of the size of the test size is small, so the inaccuracy of prediction will appear. While for test size 0.5, the difference between train means squared error and test mean squared error is very small, so the prediction is accurate.

Table 10: Train test evaluation for the linear regression for NYC Airbnb houses' price during June 2019

	Train Mean Squared Error	**Test Mean Squared Error**
0.1	4806.90411	4166.066877
0.2	4867.742578	4842.034391
0.3	4960.478117	4685.086025
0.4	4835.6305	5022.488544
0.5	4773.327776	4792.360688

4. Conclusion

4.1 Conclusion

This research studies the pricing strategy of Airbnb’s online marketplace. In this research, by employing linear regression models, time-series analysis, and using mean squared error to test the prediction accuracy, we conclude that variables without time such as house location, users’ review, availability of houses, room type, minimum nights required, and time-dependent variable like day of week, and month of year, all affect the house pricing of Airbnb to different extent. In addition, the prices exhibit obvious trend and seasonality due to weekend effect and holiday effect.

Since we use big data analysis for the house rental research, the accuracy of the prediction is high. Moreover, the dynamic price analysis combined with the mapping visualization of geofigureical information, makes our analysis very intuitive and insightful to both renters and hosts. By knowing basic information like location, number of beds/bedrooms, availability, etc., hosts have access to real-time price trends, which help them set the right price for houses and maximize their revenues (Jawad Khan, 2015). For users, simply by choosing the dates and other requirements for the house, they can find the predicted price of the house and compare prices between different hosts. Finally, it produces huge benefits for sharing economy and the society by providing valid information for future researches.

4.2 Outlook for future

Extending our research, there are various directions to pursue for future studies. Listings' substitutability and cluster analysis can be applied with aspect of house density to get finer analysis on the impact of the density of similar houses in the neighborhood on pricing—whether areas with more listings make it more competitive and lowers the average price, and how this effect varies in different clusters. It can be also combined with the map of different locations together with the existing model for dynamic pricing. In addition, customer ratings and reviews can be incorporated and analyzed using Natural Language Processing (NLP) and machine learning. NLP of the users' reviews can help us understand customer sentiment and study its impact on the house pricing as the hosts' response, and these qualitative data can be enhanced by quantitative analysis with the big data in order to elaborate the pricing prediction. Analysis and discussions about competition with similar service providers (like hotels and other platforms) will also identify the price strategies and trend for competitors that Airbnb hosts can incorporate into their pricing model.

Reference

[1] T, Lewis. "Airbnb Statistics 2019 (Growth, Revenue, Hosts More!)," May 31, 2019.

[2] Cox, Murray. "Inside Airbnb: adding data to the debate." Inside Airbnb [Internet].[cited 16 May 2017]. Available: http://insideairbnb.com (2017)

[3] Kersloot, Jan, and Tom Kauko. "Measurement of housing preferences: A comparison of research activity in the Netherlands and Finland." Nordic journal of surveying and real estate research 1, 144-163.(1994) (2004).

[4] Airbnb, "About Airbnb", *Airbnb: About | Linkedin*, Airbnb, https://www.linkedin.com/company/airbnb/about/, 11/1/2019

[5] Guttentag, Daniel. "Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts." (2016).

[6] Guttentag, Daniel. "Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector." Current issues in Tourism 18.12 (2015): 1192-1217.