The Effects of Locational and Listing Characteristics on the Price of Airbnb Listings

Justin Yu - 1006747111

29 March 2023

Figure 1: Airbnb Logo taken from website: `https://commons.wikimedia.org/wiki/File:Airbnb_Logo_B%C3%A9lo.svg`

Contents

1 Introduction

The introduction of Airbnb to the housing market in 2007 was merely spectacular compared to what it has become in the present day. Airbnb helps alleviate the stress of finding housing when going on vacation or temporary visits to places and has since become a common thing used to find good deals on housing when going out. However, where there are many people trying to help alleviate the stress coming from finding temporary housing, there are people that are trying to profit from it. Instead of Airbnb being a place where people help one another find places to live while they are temporarily away, it has become a for-profit market. People are incentivized to maximize profits over helping out one another, by for example raising the minimum nights needed to spend at listings with higher prices.

The main objective of this paper is to investigate the relationship between various attributes (such as geographic advantages and property features) associated with an Airbnb listing and its corresponding rental cost. For example, does the type of room, the minimum number of nights to stay, and the number of listings the host have an effect on an Airbnb rental price? What about the socioeconomic characteristics tied to an Airbnb's location, such as the neighbourhood crime and income levels? In analysing these relationships, we can learn more about the pricing strategies when it comes to listing a person's house on Airbnb, which can benefit both guests and hosts. For hosts, analyzing these characteristics can help them determine an appropriate rental price to charge for their property, while for guests, this paper can help them make informed decisions about what to look out for and which Airbnb listing to book.

1.1 Literature Review

Airbnb has become a popular accommodation option for travellers around the world due to its affordable prices and unique experiences. As a result, numerous studies have been conducted to investigate the factors that can influence Airbnb rental prices.

Location is one of the key determinants of Airbnb rental prices as Airbnb listings are commonly located in highly dense population sites such as urban centres within countries (Teh et al, 2020). On the other hand, properties located in less popular tourist destinations, like suburban areas tend to have lower prices. (Chica-Olmo et al, 2020) observed these effects with a hedonic spatial price model and found that walkability, noise and accessibility to urban centres are significant effects which determined the price of Airbnb apartments in Málaga (a municipality in Spain).

Besides locational factors, various listing characteristics can also influence Airbnb rental prices. These include the number of bedrooms, bathrooms, and guests a property can accommodate, as well as amenities such as Wi-Fi, parking, and air conditioning. Studies have found that properties with more rooms, and amenities tend to have higher prices (Toader et al, 2021; Teh et al, 2020).

Overall, the literature seems to suggest that Airbnb rental prices can be influenced by various factors, like location, listing characteristics like reviews and ratings, and also regulatory factors. This paper seeks to expand on these past studies by examining the complexities behind determining the rental price in large Cities like New York.

2 The Context and Data

We gathered our dataset from the data source provider, Kaggle, which provides a wide offering of datasets for data analysts. The dataset "New York City Airbnb Open Data" was chosen as it contains multiple variables on the listing characteristics of New York City Airbnb properties in the year 2019. It is important to note that the data set used does not actually represent the observed number of Airbnb properties listed over the span of 2019, but rather a snapshot of listings for a current day chosen in the year 2019. This would mean that the data set is structured cross-sectionally with no regard for time differences in comparison to analyzing panel data. Our unit of observation for the data set would be that of an NYC Airbnb Listing.

Using the geospatial variables provided in the data set such as the longitude, latitude, and Neighbourhood, we are able to combine the existing 2019 Airbnb data with other variables that we would like to analyse such as the socioeconomic conditions tied to a location. As discussed previously, we are interested in analysing whether crime can have an effect on the prices of these listings. This is because areas with high population density such as urban city centres are known to have higher crime rates and higher rental rates. However, poorer neighbourhoods may also suffer from higher crime rates but also have lower rental rates. To explore this relationship more in-depth, crime data was gathered from the New York City Police Department and was successfully merged with our current data set by aggregating the number of felonies committed near 0.2 kilometres of each listing in the year 2019.

Another socioeconomic variable which we would like to look at would be the household income levels of Airbnb hosts. Perhaps there is a relationship between high household income earners and the amount of rent that they expect guests to pay. Adding household income into our analysis would enable us to delve deeper into this relationship. However, gathering household income data for each Airbnb user is time-consuming and violates privacy laws, making it not feasible. We can solve this by looking at the median income at the neighbourhood level instead of the individual level. Median household income at the neighbourhood level is more attainable and allows for the anonymity of individuals while still providing a representative view of the population. To gather this data, HTML web scrapping was utilized, to scrape data from Statistical Atlas which provides the median income level of each neighbourhood across the United States. We limit our web crawl area to only neighbourhoods in New York City.

Lastly, tourism can be a factor that affects the price of an Airbnb listing. But tourism can be measured in various ways, for example, it can be measured by the number of visitors or by the places where the most tourist money is spent. Let's assume that we measure tourism by the most frequented place by foreign visitors. According to an article published on Wikipedia, the most visited place in New York City is Central Park followed by Times Square, and then many other attractions, all of which, garnered more than 2 million visitors in a year. We are interested in whether there is a relationship between the distance to the closest attraction and the rental price of a listing. Thus, we can use HTML web scraping to extract the names of the attractions along with the location shown in the table on Wikipedia. Using the geographical coordinates of these tourist attractions, we can measure the distance to each Airbnb listing to find and measure the distance of the closest visited attraction to the property. Finally, We can merge our newly found data on crime, household income, and distance to attraction to our current data set which already contains listing characteristics such as the room type, the minimum nights needed to be spent, and the number of listings the host has. Before any analysis can be done, data visualization techniques must be applied to the collected data.

3 Visualisations and Summary Statistics

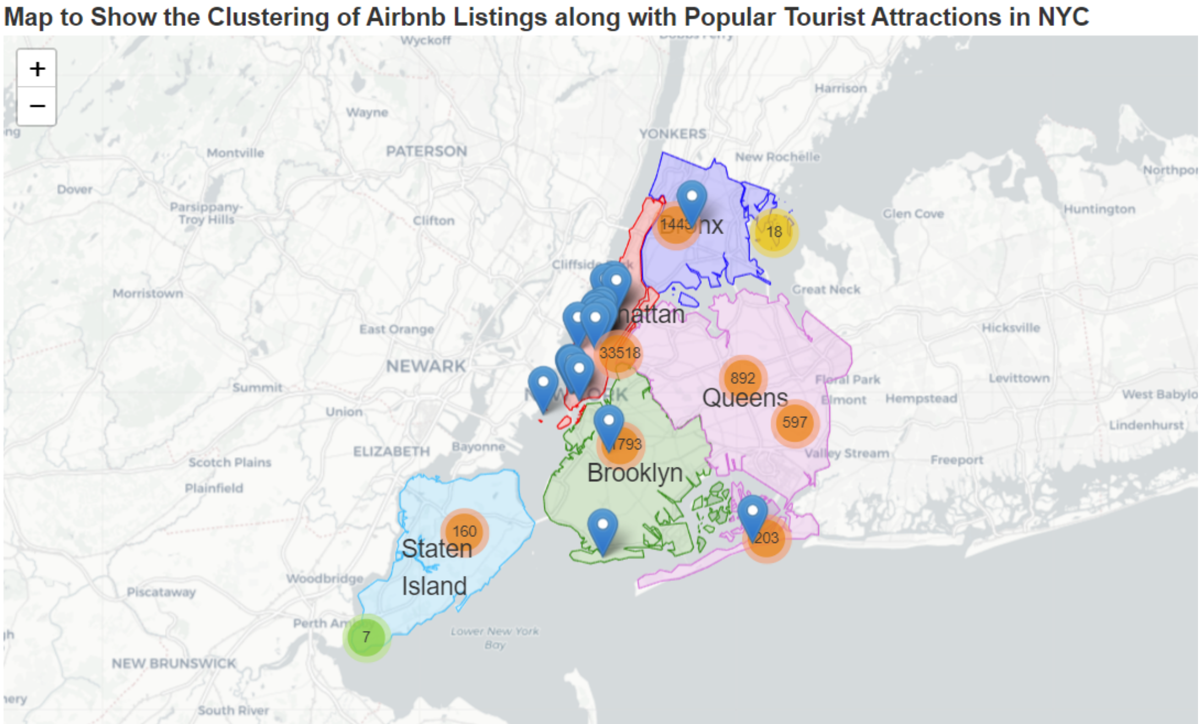

Figure 2: Airbnb Listings and Popular Tourist Attractions in NYC

Initially, we create a visualization that generalizes the geographic locations of our observations whilst marking the most visited places in New York City. From our map above we can get a general sense of the areas where users are listing their homes on Airbnb. We can see that the density of listings is high between the boroughs of Manhattan, Queens, and Brooklyn as approximately 33518 observations are clustered near the shared boundaries of the three boroughs. The Manhattan borough seems to have the most tourist attractions with around 90% of the most visited attractions being found in that borough. This makes sense as Manhattan is the most visited part of New York City, yet it has the smallest land size of the boroughs. Despite having a land size that is twice that of Manhattan, according to the cluster map, Staten Island is the least dense of the boroughs in terms of Airbnb listings. This map helps us envision where most observations are located which would be helpful in determining whether or not there are outliers in the data later on in our analysis. Our analysis of this graph can be linked back to previous studies in the past which linked an increase in the listing of Airbnb properties for more densely populated areas.

Table 1: Summary Table of the Continuous Variables in Our Analysis

	Rental Price	Minimum Nights	Host Listing Count	Distance to Closest Attraction	Household Income	Crimes Within 0.2 km
count	48631.00	48631.00	48631.00	48631.00	43517.00	48631.00
mean	141.31	6.78	7.16	3.38	70794.02	112.66
std	116.74	16.12	33.03	2.59	32087.06	78.01
min	10.00	1.00	1.00	0.01	22652.04	0.00
25%	69.00	1.00	1.00	1.46	46899.14	55.00
50%	105.00	3.00	1.00	2.85	54422.66	98.00
75%	175.00	5.00	2.00	4.61	91279.06	151.00
max	1000.00	365.00	327.00	23.70	225119.82	914.00

The summary table presented above provides a concise overview of the continuous numerical variables that will be utilized in our analysis. Our data contains a total of 48,631 observations for almost all of the variables except Household Income (which has 43,517 observations). This difference in around 5000 missing observations presents itself as a problem as running a regression with missing values for some predictors and not others, can lead to misleading or inaccurate estimates. Therefore, it is important to handle this missing data issue appropriately by either imputing the missing values or excluding the incomplete cases from the analysis. Since we are already dealing with such a large sample of data, we choose to omit these missing values from our models.

Measurement errors should be removed from our data set as it affects the precision of analysis, however, determining what is a measurement error can be difficult. Since the rental price is chosen by the host, it is unclear whether or not the price chosen was an input error or a clear judgment made by the host. Thus, to mitigate such extreme outliers from affecting our analysis, the rental price was capped at \$1000 per day reflected in the table. We can then use the summary table to help visualize the distributions of each variable. For the dependent variable Rental Price, a mean of \$141.31 and a median of \$105.00 was obtained. When the mean is greater than the median, this usually indicates that the distribution of the variable is positively skewed as the mean is pulled higher due to extreme values. Capping the rental price to \$1000 only removed some of the extreme values from our price variable, and we can see that there still exist large outliers in the data. Removing the rest of these extreme values would decrease the standard deviation and the skewness but it also removes the reliability of the estimates generated since the observations are not errors and are a part of our population.

Using this theory to gauge our independent variables, we observe that most variables do not appear to be normally distributed, where in that case, the mean and median would be roughly the same. The Household Income Variable has a large standard deviation (\$32087.06) which means large deviations from our sample mean. We can adjust this by scaling down the variable by measuring income per $10000 instead of per $.

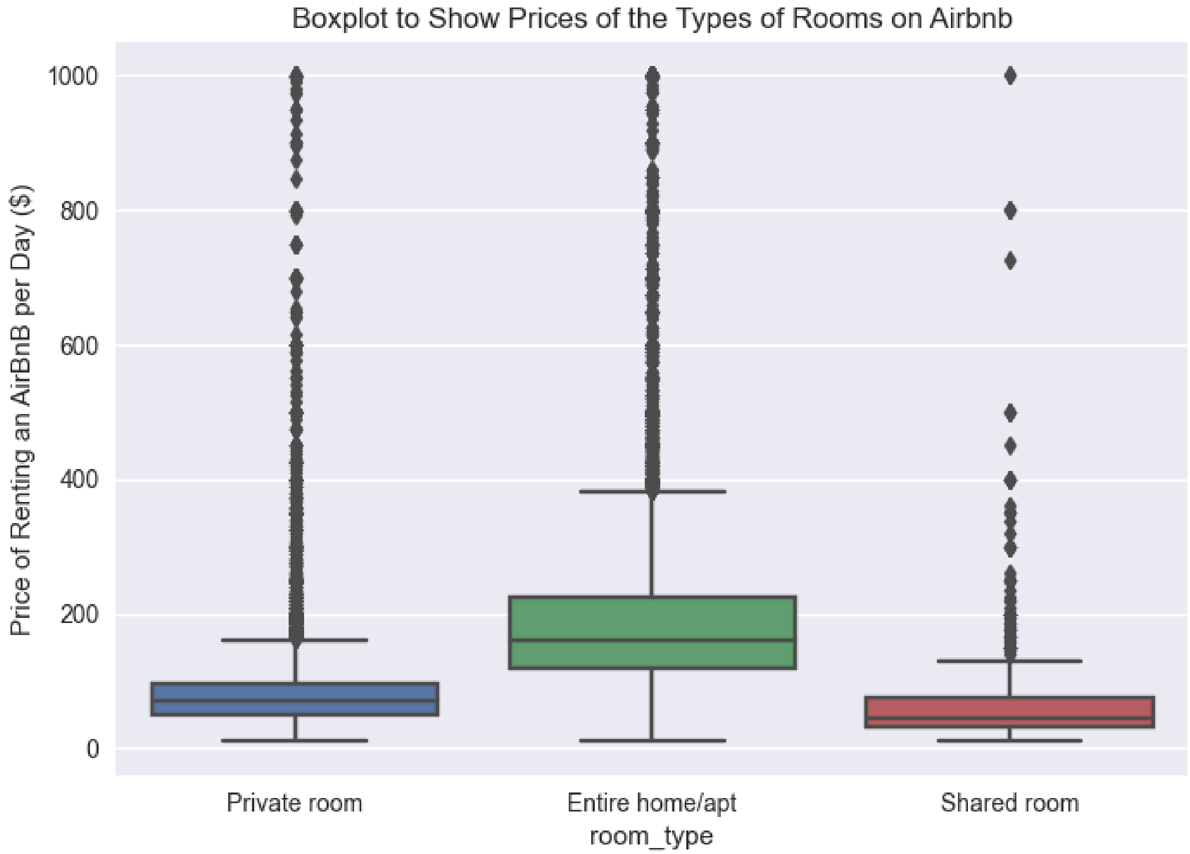

Figure 3: Prices of the Types of Rooms on Airbnb

Having examined our numerical variables, we will now shift our focus to the categorical variables used in our analysis. From the Boxplot shown above, we look at the relationship between the rental price and the type of room being offered in the listing. We can see that the rental price of entire properties is noticeably higher than that of private or shared rooms with a median of approximately $180 per day. Comparing private room and shared room listings, we can see that the price for both categories are relatively the same with the median of shared room listings being lower by a bit. The box plot also highlights again, the problem of large outliers. This is seen by the set of observations that lie 1.5 times the Interquartile Range above the third quartile (75%).

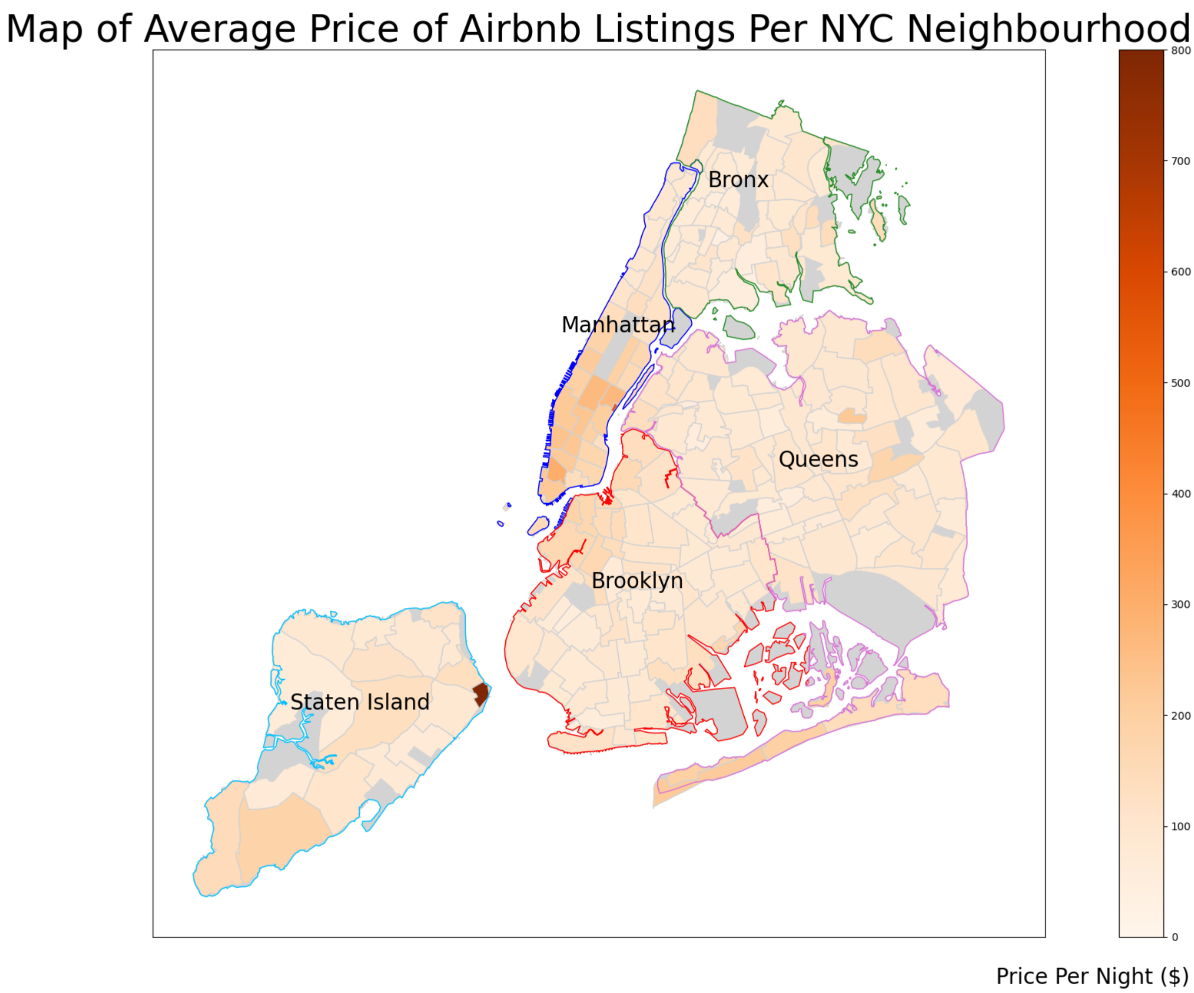

Figure 4: Airbnb Rental Price Averaged at the Neighbourhood Level

The choropleth map above shows the average Airbnb rental price for a neighbourhood in each of the five boroughs in New York City. From the map, we can see that the average Airbnb listing is usually around the 100 - 400 mark in each neighbourhood with generally higher prices going closer to the center of the city. Taking a look at Manhattan and comparing it to its neighbouring boroughs, we can see a general darker shade of around 300 - 400 dollars on average for listings in its neighbourhoods whilst the neighbourhoods in the boroughs Brooklyn and Queens that are near Manhattan are in the 200 to 300 dollar range. We find that most of the points shaded in darker orange that is away from Manhattan are found on the coastlines of the boroughs like Staten Island and Queens. This seems to make sense as lakefront properties usually are priced higher than inland properties based on the promise of being close to the beach.

Overall, the average price of Airbnb listings for each New York neighbourhood seems to not be drastically different with an average difference of $100 between higher and lower-priced neighbourhoods. We do observe an outlier such as one of the neighbourhoods in Staten Island which has an average neighbourhood listing price of $800 per day. However, taking a look again at the first Map, we can see that the count of listings drops when moving further away from the intersection of Brooklyn, Manhattan, and Queens. Using this information and the average prices per neighborhood in this map, we can deduce that there may not be enough observations in the neighbourhoods as we move away from Manhattan thus giving a bigger weight to the more expensive listing to skew the average of the house prices in those locations.

But what does this mean? We know that the number of Airbnb listings increases in areas with higher population density and more tourist attractions (Figure 2). However, this does not mean inherently that the overall price of Airbnb listings increases else we would have seen a large difference in the neighbourhood prices from comparing boroughs like Manhattan and Staten Island. This implies that in these same densely populated areas, there are just as many offerings of affordable temporary housing as there are for expensive properties. Thus the average rental price of these neighbourhoods in Manhattan is pulled downwards due to the offering of affordable properties.

But what are the benefits for hosts pricing their properties at a lower cost in higher tourism areas? Perhaps there are some good people out there providing lower-cost housing at their expense. To be more realistic, the reason why hosts price their properties lower is that they are able to make more profit, due to the high demand for their property. For example, a host which prices their unit for a lesser price can make more money due to the higher demand than if they had priced it at a higher value and the unit is left empty for longer periods of time. This method of lowering prices for higher demand and possibly larger profits provides incentives for hosts to lower the price of listings. But how would this apply to less touristic areas like Staten Island? This ties back to the demand and supply of housing as unlike Manhattan there is a low demand for temporary housing in Staten Island. The reason for this low demand can be due to its proximity to its lack of tourist attractions. From Figure 2, we see that there are no widely visited attractions in Staten Island. With tourists making up a large percentage of Airbnb users, the lack of tourist attractions means a lower demand for temporary housing in the borough. Because of this, the method of lowering prices for higher rental rates is unrealistic as there are already low rental rates in that area. Hosts would need to adapt to this different market as temporary housing demand in those areas is not determined largely by tourism but by other reasons like visiting family.

4 Results

Table 2: Regression Analysis of the Rental Price of Airbnb Listings

	(1)	(2)	(3)	(4)
Dependent Variable:	Price	Price	Price	Price
Host Listing Count	0.130***	0.123***	0.099***	0.099***
	(0.015)	(0.015)	(0.015)	(0.036)
Minimum Nights	-0.370***	-0.370***	-0.374***	-0.383***
	(0.030)	(0.030)	(0.030)	(0.044)
Entire	109.430***	109.730***	110.041***	109.549***
	(3.139)	(3.138)	(3.131)	(4.879)
Private	17.915***	18.171***	18.871***	18.484***
	(3.135)	(3.134)	(3.129)	(4.372)
Distance to Closest Attraction	-3.423***	-3.192***		
	(0.218)	(0.221)		
Median Income	0.001***			
	(0.000)			
Crimes Within 0.2 km	0.118***			
	(0.007)			
Inverse of Distance		2.698***		
		(0.437)		
Log of Distance			-16.602***	-13.319**
			(0.807)	(5.222)
Income Measured in 10k		7.295***	6.178***	5.779***
		(0.171)	(0.192)	(1.417)
Crime Per Month		1.364***	1.222***	0.683*
		(0.083)	(0.084)	(0.351)
const	21.467***	21.049***	36.524***	27.354*
	(3.610)	(3.609)	(3.774)	(16.456)
Controls	No	No	No	Yes
Robust	No	No	No	Yes
Observations	43,508	43,508	43,508	43,508
R^2	0.304	0.305	0.307	0.309
Adjusted R^2	0.304	0.305	0.307	0.309
Residual Std. Error	96.429(df = 43500)	96.388(df = 43499)	96.234(df = 43500)	96.073(df = 43496)
F Statistic	2715.260*** (df = 7.0; 43500.0)	2382.654*** (df = 8.0; 43499.0)	2751.463*** (df = 7.0; 43500.0)	165.401*** (df = 11.0; 43496.0)

Note: *p<0.1; **p<0.05; ***p<0.01
Binary variables were added to control for differences seen across different Boroughs

We generated 6 different models for our regression analysis, each with different functional specifications. Model (1) first looks at an assumed linear relationship between each regressor and the rental price. Statistically significant coefficients were obtained for all the independent variables and can be attributed to our large sample size. This means that there is evidence of a relationship between each independent with the dependent variable, and that this relationship is unlikely to be due to chance. For that model, we obtained a R^2 of 0.304 indicating that roughly 30 percent of the variation of Airbnb rental Price is determined by the independent variables. However, a purely linear model is likely not representative of the true functional form of the relationship. Observing the relationship between the independent variables and the dependent variable from Figure 5 in the Appendix, the relationship appears to be non-linear.

We know that violations in a regression functional form lead to biased estimates so to mitigate this issue, we attempt to fit a non-linear multiple regression model shown in column (2). For this model, we added a polynomial term related to the distance of an Airbnb listing to a tourist attraction. Household Median Income and Crimes committed near the property were also scaled down to determine a better relationship as a \$1 change in neighbourhood median household income led to a change in the price of an Airbnb listing by \$0.001. Model 2 fixes these issues as the relationship is better seen and the R^2 increased from 0.304 to 0.305 indicating a slightly better fit.

Model 3 takes it a bit further by looking at the linear-log relationship between the rental price and the natural log distance to a tourist attraction. Fitting the model with the natural log distance to a tourist attraction increases the R^2 of the model to 0.307. The relationship between Price and Distance can now be interpreted as a 1% change in the distance (in km) to a tourist attraction leading to a decrease in the rental price (Y) by (0.01 X -16.60 = -\$0.17). We can also add binary variables which indicate which borough, the listing is found in. The addition of these binary variables in column (4) helps us to control for unobserved differences in the boroughs. Lastly, we cluster our standard errors at the neighbourhood level to account for potential correlation among observations within the same neighbourhood. In doing so, we found that our standard errors for each coefficient increased, which leads to smaller t-values and thus larger p-values. Most predictors remain significant at the 1% significance level, however, crime per month and log distance to a tourist attraction does not seem to be significant at the 1% level anymore.

Table 3: Regression Analysis of the Log Rental Price of Airbnb Listings

	(1)	(2)
Dependent Variable:	Log Price	Log Price
Host Listing Count	0.000***	0.000*
	(0.000)	(0.000)
Minimum Nights	-0.003***	-0.003***
	(0.000)	(0.000)
Entire	1.091***	1.091***
	(0.015)	(0.035)
Private	0.364***	0.364***
	(0.015)	(0.030)
Log of Distance	-0.084***	-0.084***
	(0.004)	(0.028)
Income Measured in 10k	0.032***	0.032***
	(0.001)	(0.007)
Crime Per Month	0.004***	0.004**
	(0.000)	(0.002)
const	3.639***	3.639***
	(0.033)	(0.104)
Controls	Yes	Yes
Robust	No	Yes
Observations	43,508	43,508
R^2	0.521	0.521
Adjusted R^2	0.521	0.521
Residual Std. Error	0.462(df = 43496)	0.462(df = 43496)
F Statistic	4301.285*** (df = 11.0; 43496.0)	310.688*** (df = 11.0; 43496.0)

Note: *p<0.1; **p<0.05; ***p<0.01

Binary variables were added to control for differences seen across different Boroughs

Instead of looking at a linear dependent variable, we can change our model to instead analyze the effects of the predictors on the log of the rental price of a listing. Fitting this regression model in column (1) we get a R^2 value of 0.521 meaning that 52.1% of the variation of log rental price is explained by the model. We see that the slope coefficient of our predictor variables falls in the log model, but this does not mean a lesser effect, but rather a difference in interpreting the values. For example, looking at the relationship between Entire and Log Price we can see that being an entire property increases the rental price by like (100 X 1.091 = 109)% as opposed to not being an entire property. We can finalize this model by clustering it at the neighbourhood level as we did for our Price model. In doing so we found that the predictors, host listing count, and crimes per month are no longer significant at the 1% level. We chose our best model which measured the relationship between Rental Price an the predictor variables to be Table 2, column (2). This means that our regression equation would measure:

$$\log(RentalPrice) = \beta_0 + \beta_1 HostListingCount + ... + \beta_7 CrimePerMonth + \eta_i + \epsilon \quad (1)$$

where:

β_0 and β_1 are called parameters or coefficients

ϵ is the error term

η_i is the set of dummy variable controls related to the effect different boroughs have on the rental price

Model 6 tells us a multitude of things about the relationships between the predictor variables and Price. Going down the list, we can see that hosts having more listings on Airbnb do not seem to have a percentage increase in the price of an Airbnb listing. This predictor also lacks in significance given that it is only significant at the 10% level. We can see that an increase in the minimum amount of nights required to stay by one unit, leads to a decrease in the rental price by 0.3%. This may not seem significant, however, know that for an increase in the minimum nights required to stay by one day for a listing priced at 100 dollars, this decreases to 70 dollars, holding the other characteristics constant. Crime per month on the other hand, seems to increase the rental price by 0.4%. This relationship however is only significant at the 5 percent levels and raises some doubts about the predictor's relationship with price.

5 Limitations

The analysis presented in this study is subject to several internal and external validity issues that should be taken into consideration when interpreting the findings.

Externally: While New York City is one of the Largest Cities in the World with 8.46 million residents, there may lie differences between cities which make generalising these results to other cities difficult. As mentioned in previous studies, rental prices can be influenced by regulation (Toader et al, 2021). Thus differences in the types of short-term rental regulations would make generalising this study to other areas in the world or even in the United States difficult.

Internally: Perhaps some hosts undervalue their homes and some hosts overvalue their homes, not knowing their true worth. This randomness of pricing by an individual can explain some of the non-linearity observed in rental prices and the predictors. Another factor that can contribute to non-linearity in rental prices is the fact that the rental price of listings is chosen by individuals, and these individuals may have different motivations for setting the price of their property. For example, some hosts may be more focused on maximizing their rental income, while others may prioritize filling their property with tenants quickly. The complexities around humans choosing the price of their rental can lead to biased estimates and a nonlinear functional form. Even if we were to fit a perfect nonlinear model, interpreting the coefficients of the model would be difficult.

6 Conclusion

This paper aimed to explore in-depth, the relationship between multiple locational and listing characteristics on the rental price of an Airbnb Property.

Our analysis has shown that location plays a significant role in determining rental prices, with properties located in highly dense population sites such as urban centres having higher prices. Additionally, listing characteristics such as the minimum nights required to be spent and the type of room being listed, also influence rental prices. With further testing, we discovered that the number of listings owned by the host has a very small effect on the prices of Airbnb Listings.

From the maps provided, we can see that there seems to be a lack of observations of listings as we move away from these listings. This reduction in listings meant that listings with higher prices have a larger weight on the average Airbnb price in that neighbourhood. We discovered that this can be because there are fewer incentives for listing homes on Airbnb provided that it is not in tourist locations like Manhatten. Listings which are found away from tourist locations are more likely to be higher in price as there seems to be a luxury component associated with the listing as producers adapt to the demand preferences of consumers. We can say that since people are less likely to demand temporary housing away from high tourist destinations, producers must find a different way of incentivising their products in order to make a profit. In combining our listings with crime data we can see that the prices of Listings do increase with the level of felonies in that area from our regression model. This relationship is only significant at the 5% level implying that the relationship may be false. Perhaps comparing Airbnb listing prices to the level of felonies per month is not a viable option.

Scraping data from the internet we were able to find and combine our current data with the median household income across different neighbourhoods in New York City. In analysing the relationship between income and rental prices, we do see an increase in the rental price in areas where there are higher-income people. This, however, does not entirely explain the relation behind why prices are high, but does seem to present a case where neighbourhoods with higher median household income demand higher rents or offer higher-end properties. But we also see that in these same high-income neighbourhoods there, are supplies of affordable housing which decreases the overall average cost. Whereas locations where there is less tourism have lower maximum prices charged but also higher minimum costs which raises the average costs of housing.

Lastly, we have identified measurement errors as a challenge in our analysis, but we have mitigated this by capping the rental price at $1000 per day. Rental price determination is very complex and further studies need to be conducted in order to strengthen this paper's analysis. The covid-19 pandemic sent shocks throughout the entire Airbnb market affecting the incomes of many people. Perhaps a difference in difference model could be at play here.

7 References

Chica-Olmo, J., González-Morales, J. G., Zafra-Gómez, J. L. (2020). Effects of location on Airbnb apartment pricing in málaga. Tourism Management, 77, 103981. https://doi.org/10.1016/j.tourman.2019.103981

Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Current Issues in Tourism, 18(12), 1192-1217. https://doi.org/10.1080/13683500.2013.827159

Teh, P. L., Low, Y. C., Ooi, P. B. (2020). Viewing airbnb from Twitter. Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications Services. https://doi.org/10.1145/3428757.3429112

Toader, V., Negrușa, A. L., Bode, O. R., Rus, R. V. (2021). Analysis of price determinants in the case of Airbnb listings. Economic Research-Ekonomska Istraživanja, 35(1), 2493–2509. https://doi.org/10.1080/1331677x.2021.1962380

Datasets:

DGOMONOV. (2022). [New York City Airbnb Open Data], [Version 3]. Retrieved [03/03/2023] from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

https://www.kaggle.com/datasets/brunacmendes/nypd-complaint-data-historic-20062019

https://statisticalatlas.com/place/New-York/New-York/Household-Income

8 Appendix

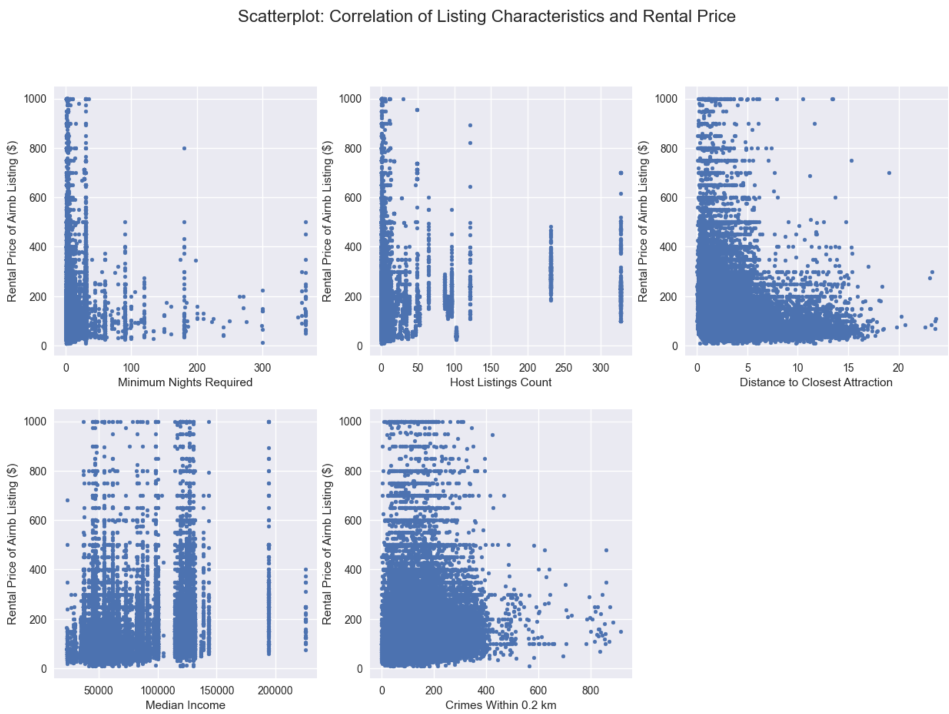

Figure 5: Correlation of Listing Characteristics and Rental Price