1. Introduction

Airbnb is a multi-sided online platform for listing and renting local homes, connecting hosts and travelers from all around the world. The sharing economy allows resources to be shared and used without possessing them, providing travelers the flexibility in choosing their preferred accommodations from a wide variety in the vicinity, and hosts the opportunity to list their spaces whenever available, and generate income from rental revenues. Founded in 2008, Airbnb has more than 4 million listings (Mody, 2018), infiltrating more than 65,000 cities and 191 countries (Bivens, 2019). It became the leading lodging company globally as of November 2016, by the number of accommodation listings (Haywood, 2016). For the purpose of our project, our group will specifically be looking into accommodation listings – such as homes, cottages, shared rooms, individual rooms which are the primary service provided by Airbnb.

Airbnb can be perceived as a competent substitute for the services offered by the hotel industry. Positioning itself to be a home away from home option, Airbnb facilitates access to local culture, activities, and people, boosting tourism especially in underexplored areas in cities. Guests gain a cheaper, more authentic travel experience, as compared to conventional hotel accommodations that are more expensive and rigid in terms of customization. Unlike mainstream hotels and hostels, Airbnb provides a plethora of accommodation choices to suit one's needs and preferences, since each accommodation listing is different from another. Outdoor glamping, treehouses, and houseboats are some unique types of accommodations Airbnb offer. The Airbnb platform is designed in a manner that gives users the feeling that the service was created for them. The clean-design and intuitive website and mobile platforms emphasise on personalisation. Thus, Airbnb's entrance into the hospitality industry leads to much attention to the disruptive nature of the service.

We have analysed prior research works that provided us with valuable insights towards Airbnb's impact on hotel revenues, breaking down the general aspects of the Airbnb experience. Airbnb accommodation characteristics that influence the demand for a particular listing were evaluated for their level of significance, such as value-to-price, and review ratings. However, insufficient research has been done towards the extensive characteristics of Airbnb listings and reviews, from the user experience of browsing for accommodations to a successful homestay. These research studies view NYC's rental prices as homogenous throughout the city, and are not controlled for the characteristics of different neighbourhoods within NYC, such as the disparity in socio-economic levels and rental prices that exist among neighbourhoods in NYC. With these, the motivation of our project is to delve into a more extensive evaluation of Airbnb's listings and reviews. Our study attempts to understand the complex and extensive relationships and level of influence of different Airbnb features that customers look out for, in the hopes that hosts can leverage on these insights to increase occupancy rates of their listings and generate higher revenues.

The objectives of this project are to carry out both demand side and supply side analysis by assessing factors that impact yearly occupancy rate and hosts set prices in the Airbnb landscape. This allows us to generate insights on the features that are significant in influencing a guest's decision to book a particular listing and also the ones that are influential when hosts determine the price for the listings. In particular, we would like to investigate the impacts of the focal variables: 1) inter-market competition level, 2) product attributes of the listings, 3) consumers' perceptions in determining the demand and supply outcomes. Occupancy rate is used to represent the demand outcome and listing price is used to represent the supply outcomes. We aim to help hosts and Airbnb marketers looking to improve their overall occupancy rates and increase rental revenues from Airbnb.

We focused our study on New York City's Airbnb market. New York City is the third-largest Airbnb market globally and a popular tourist destination. More than 15% of accommodations in NYC are Airbnb listings (Roach, 2018). The mature Airbnb business in NYC provides a rich and sufficient dataset that captures most of the consumers and hosts' behaviours, leading to more reliable findings. Furthermore, other relevant information about NYC's public transportation and major attractions are more accessible online more comprehensive complementary information to work with. Based on the datasets retrieved from InsiderAirbnb (listings, reviews, and calendar) complemented with NYC's public transportation data, we generated a panel data from 2016-2018. Relevant variables are generated through feature engineering using text analysis, image properties analysis and descriptive analysis. Our econometric specification models both listings' occupancy rate and listings' price as a function of listings' competition, listings' attributes and consumers' perceptions, controlling for relevant factors to carry out demand-side analysis and supply side analysis respectively. Our identification strategy employed linear regression using clustered standard error, fixed effects model and random effects model to quantify the impacts of the focal variables on the demand-side outcome which is measured by occupancy rate, and the supply-side outcome which is measured by price.

We find evidence that the above mentioned focal variables play an important role in determining the occupancy rate and prices. The effects of variables differ in determining the demand which is shown via the occupancy rate model and the supply which is shown via the price model. Inter-market competition provides a wide range of substitutes for the consumers, thus decreasing both the occupancy rate and the ability to raise prices. One percent increase in houses count in the same neighbourhood will lead to 0.015% drop in occupancy rate and 0.016% drop in listing prices. While different types of amenities play different roles in determining the supply and demand outcomes. It offers insights that not all amenities might be desired. Among all, child-friendly amenities play the most crucial in determining both occupancy rate and listing prices. 1 unit increase in the number of child-friendly amenities variable is expected to lead to an increase of 149.8% in listing price and 0.109 increase in occupancy rate. Furthermore, consumers' reviews are proven to be a key determining factor when future consumers are making their purchase decisions. Greater number of reviews and more positive reviews increase demand and also enhance hosts' ability to raise prices. This highlights the importance of the Peer-to-Peer review model that Airbnb has adopted. The findings from our project offer meaningful business insights for Airbnb marketers and hosts to boost their future revenues and attractiveness in the market.

2. Literature Review

Previous research has been invested in Airbnb's effects on related industries, as well as its characteristics, such as user-generated ratings. These are especially relevant to us as these are the features that we are to include in our model.

Relevant research has been carried out to investigate the impact of Airbnb on the hotel industry which can be perceived as a close substitute for each other. One research paper (Zervas and Byers, 2014) quantifies the impact of Airbnb on the hotel industry – the extent to which Airbnb's growth serves as a substitute for hotel accommodations, and its impact on the revenues earned by affected hotels. In this research, a dataset spanning all Airbnb listings in Texas, as well as a 10-year panel data of quarterly tax revenues for all Texas hotels were collected. A difference-in-difference analysis and linear regression model were employed to develop an estimate of Airbnb's impact on hotel revenues. Hotel segments that consumers are less likely to substitute for Airbnb stays were added as control groups. It is found that a 1% increase in the number of Airbnb listings in Texas results in a 0.05% decrease in quarterly hotel revenues. Also, the impacts of Airbnb are found to be distributed unevenly across the industry, disproportionately impacting lower-end hotels, as one might expect given the similar nature of rentals on Airbnb, typically with fewer amenities and services. This paper concedes the endogeneity Airbnb market entrance faces – the time-varying unobserved factors that are correlated to Airbnb adoption and hotel revenues. However, the authors posit that Airbnb adoption by hosts is driven opportunistically by self-interest, rather than a planned business strategy against corporate hotel chains. In other words, it is highly unlikely that individuals consider hotel revenues before listing their properties on the platform. This paper asserts that the positives of Airbnb's growth must be evaluated against the current costs that hotels pay, as estimated in their paper. Our group would like to point out that the impacts of Airbnb on hotel revenues vary significantly according to the cultural environment and legislation of individual cities around the world. With that in mind, we must extrapolate the insights gathered from the Airbnb listings in Texas with strong discretion and caution, as it is not entirely representative of all cities. For example, the GDP per capita of Texas and NYC visibly differ, which reflects the difference in socio-economic levels, purchasing power, and cost of living in both cities. These are factors, not to mention different legislation laws, that may affect the demand and supply of Airbnb listings.

The next research paper (Blal et al., 2018) we dissected also analyses the impact of Airbnb on the hospitality industry – how (1) Airbnb accommodations supply (volume of listings), (2) prices of Airbnb properties, and (3) perceived quality (users' satisfaction) affect hotel sales performance. This study measures the trend of change in hotel sales performance as Airbnb's supply increased in the city of San Francisco. In this research study, the anonymized data on 101 hotels' total revenues, total supply, number of rooms were collected longitudinally. This model uses fixed and random linear and polynomial two-level models, stratified according to the 10 neighbourhoods in San Francisco. Fixed effects tested for in the research study include Unemployment rate at the time of measurement of hotel performance, dummy variables to capture seasonality (high and low seasons) and weekend effects (weekday and weekend), which can affect the RevPAR of hotels. Unlike the previous paper whose evidence that an increase in the number of listings results in a decrease in hotel revenues suggest a substitution relationship between Airbnb listings and hotel accommodations, this paper has found that Airbnb supply does not affect hotel demand and sales, indicating a supplementary effect, and in fact, the average price of Airbnb listings affect hotel RevPAR positively. The higher the price of rentals posted on Airbnb, the higher the RevPAR of hotels, and hence sales performance of hotels, since the study mentions that hotel demand is not impacted by Airbnb supply. It is also found that an increase in quality of Airbnb listing has a direct adverse impact on hotel sales performance; with every unit increase in the review score rating of an Airbnb property, there is a decrease in average hotel RevPAR of $25.54 for hotels in the data sample used. This study concludes that the impact of Airbnb is not so much influenced by the volume offered on the platform, but rather the pricing and perceived price-to-value ratio. Pricing and perceived price-to-value ratio are relevant to our project as we are looking at characteristics of Airbnb listings that guests evaluate before booking an Airbnb accommodation. Comparing this research study from the previous aforementioned, the disparity in conclusions drawn can be attributed to the fact that this study delves deeper into the characteristics of Airbnb listings – the volume of listings, prices of listings, ratings of listings, instead of analyzing aggregated Airbnb volume.

The third research paper (Proserpio, 2015) works on understanding and interpreting the nuances of user-generated ratings in the context of Airbnb as a sharing economy. With over 95% of Airbnb properties with 4.5 stars and more, there is a stark difference between Airbnb's ratings (mean: 4.7 stars) and that of TripAdvisor's (mean: 3.8 stars). TripAdvisor is the main comparison used in this research study for its worldwide scope and scale, and accommodation diversity. 412,223 hotels and 54,008 vacation rentals accommodation data from TripAdvisor, as well as 226,594 Airbnb properties, were collected, and these listings from both platforms have at least 3 reviews on their respective sites. Product heterogeneity was considered; this study examines the differences in ratings based on the type of accommodation and geographical location, since average ratings are significantly influenced by the product mix. This study then compares cross-listed properties, which are properties that have been lifted on both Airbnb and TripAdvisor. The results reveal that Airbnb ratings of these cross-listed properties are higher than their TripAdvisor ratings. This indicates that property heterogeneity alone is unable to account for the rating gap between both platforms. The potential reason suggested by the author was the observed bias arising from bilateral reviewing on Airbnb, where hosts and guests leave ratings for each other for each successful stay. Lastly, the study develops a predictive model to use Airbnb ratings to predict TripAdvisor ratings. While there exists significant positive association, the adjusted R^2 is low, therefore suggesting that the ratings on one platform explain only a small degree of variation in ratings on the other. There is a weak positive correlation between the two sets of ratings, suggesting that both platforms have distinctive preferences in ranking and rating accommodations. This research paper is useful in explaining our NYC data, where the review score ratings are mostly 90% and above. With these insights about ratings, we can measure our model results to check if the estimators corresponding to review score ratings align with this research study.

3. Data Description

3.1 Inside Airbnb Dataset

Inside Airbnb[1] is an open independent and non-commercial set of data that can be used for the analysis of Airbnb listings in various cities. The data is collected from publicly available information on the Airbnb site, and is separated by city. The scope of our project will be restricted to the New York dataset, due to the availability of complementary data to support our analysis. However, our feature engineering and research methodology can be readily applied to data from other countries given access to similar complementary data.

The Inside Airbnb dataset for New York is split into 4 files:

1. *Listings.csv* contains detailed public information regarding each listing. There are 106 columns, each containing different variables describing the listings including location, description, availability, and price etc.
2. *Reviews.csv* contains all reviews collected for each listing.
3. *Calendar.csv* contains information on each listing's availability for the next 365 days.

[1] http://insideAirbnb.com/about.html

4. *Neighbourhood.csv* contains information on the neighbourhoods and sub-neighbourhoods for the given c*ity*.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Reviews	241	1287	4656	10344	21430	43155	90280	167963	265989	439204

Additionally, we only considered data from the years 2016-2018, since this period coincided with significant growth in the number of reviews, allowing us to efficiently capture a majority of the sample, while ensuring the relevance of our results on the most recent data. Listings with no reviews over this specified period were not included in our final dataset, resulting in the following breakdown in our final dataset.

Table 3.1 Listings breakdown by year

	Listings
2016	12320
2017	17141
2018	25015

3.2 Exploratory Data Analysis

As part of our preliminary data analysis, we looked into the review data in order to identify important factors that were frequently mentioned by guests. This provides some insight into features that were highly desired by the guests, and therefore frequently mentioned, as well as insights into specific features that will be helpful in our model. The data analysis for reviews was split into 3 sections:

1. *Nouns* - Nouns generally refer to objects and places. This would track guest preferences for amenity and location features.
2. *Adjectives* - Adjectives generally refer to descriptive terms that can be used to describe the listings. This can allow us to get an idea of the type of accommodations that the guests prefer
3. *Verbs* - Verbs generally refer to action and activities, and can allow us to track what type of activities guests tend to prioritize.

The text classification was performed by NLTK's Part-of-Speech tagger, a pre-trained package for pos-tagging. The specific tags selected for each category of words were defined as shown:

Category	**Tags**
Noun	*VB, VBD, VBG, VBN, VBP, VBZ*
Verb	*NN, NNS, NNPS, NNP*
Adjective	*JJ, JJP, JJS*

Detailed descriptions of each tag can be found in the appendix.

3.2.1 Common Nouns

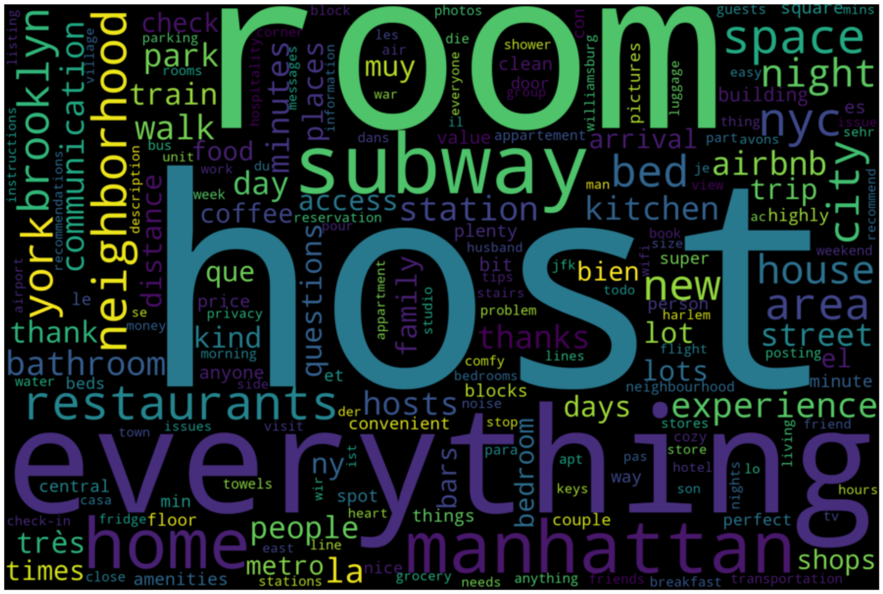
Table 3.2.1 Top noun frequencies in reviews

Neutral terms removed**:** *place, apartment, location, stay, time, great*

	Frequency
host	213262
room	158252
everything	126688
subway	123519
manhattan	93327
home	87179
restaurants	83396
space	83246
neighbourhood	81320
nyc	80526

The popularity of the term *"host"* in reviews suggest that the host plays an important part affecting guest sentiments. While the public data provided does not provide detailed information about the host-guest interaction, we can attempt to include this factor in our models by using host experience. This can be measured by either taking the age of the host on Airbnb, or the Superhost title, a title given by Airbnb to experienced hosts.

Notable factors mentioned frequently were *"subway"*, suggesting access to public transport as a significant factor, as well as location-related terms such as *"manhattan"* and *"neighbourhood"*, pointing to the immediate neighbourhood of the listing as a significant factor in determining guest satisfaction. The term *"everything"* can be ambiguous, but can also suggest that in terms of amenities and features of a listing, guests are likely to value raw quantity of amenities and features over the availability of specific amenities and features.



The word cloud for nouns shows that in addition to the top 10 most frequent words, amenities and features such as *"bathroom"* and *"kitchen"* tend to be frequently mentioned in reviews. In addition, *"walk"*, *"park"*, *"metro"*, *"station"* and *"access"* also highlight the importance of accessibility of public transport, while justifying our use of Google Maps walking distance for our distances.

3.2.2 Common Adjectives

Table 3.2.2 Top adjective frequencies in reviews

Neutral terms removed**:** *great, perfect, super, wonderful, nice, good, amazing, est, un, nan, excellent*

	Frequency
clean	222128
comfortable	140362
easy	112848
quiet	71874
helpful	71227
friendly	53987
spacious	50536
safe	46151
little	42789
beautiful	39630

For descriptive terms, the review analysis shows that cleanliness was valued by guests over all other terms. Notably, the importance of guest-host interactions is also seen here with the prevalence of human descriptors, *"helpful"* and *"friendly"*.

Most of the other high frequency adjectives tend to be neutral terms from which no meaningful or actionable insights can be drawn for our model, especially considering the dichotomy between terms such as *"spacious"* and *"little"*. The key notable words in the cloud are *"convenient"*, *"communicative"* and *"responsive"*, which again refers us back to the accessibility of the place, as well as host-guest interactions as key factors.

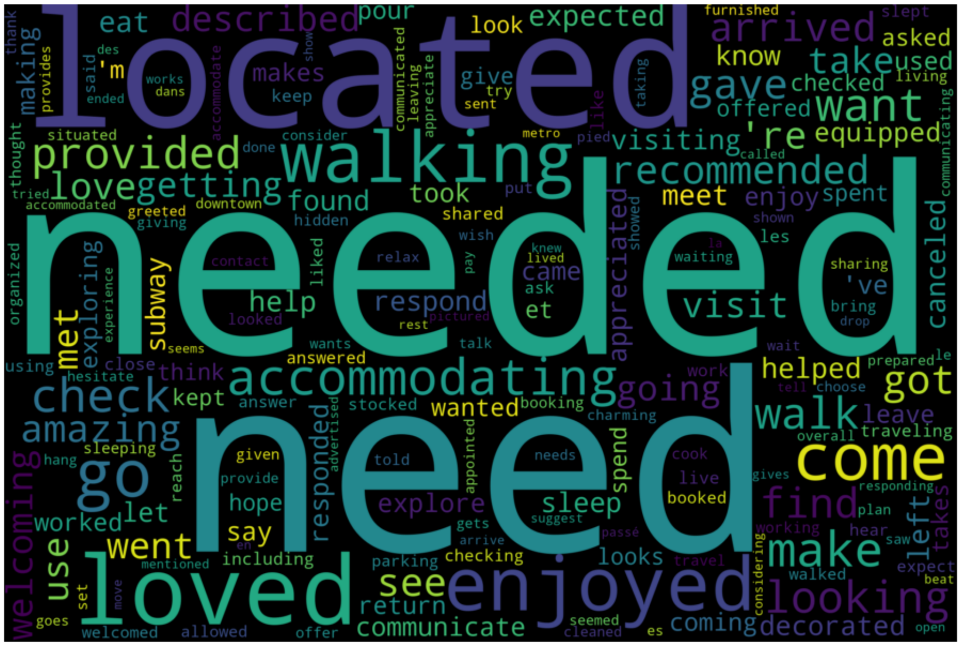
3.2.3 Common Verbs

Table 3.2.2 Top verb frequencies in reviews

Neutral terms removed**:** *great, perfect, super, wonderful, nice, good, amazing, est, un, nan, excellent*

	Frequency
needed	57906
need	53111
located	51329
loved	43909
enjoyed	42807
walking	41440
go	35561
come	34979
accommodating	33383
make	31197

Verb frequencies of *"needed"* and *"need"* suggest the significance of amenities in determining guest satisfaction, and therefore identifies available amenities as a possible independent variable for model. Notably, the term frequency of *"accommodating"* points again to the significance of the host, while the term *"walking"* and *"located"* points to the importance of specific locations identified in the noun frequency being within walking distance. This further identifies the metrics we use to quantify distance to the nearest subway stations, as well as the need to include provided amenities in our model.



Other less frequent, yet actionable terms that appear in the word cloud include *"provided"*, which suggests amenities as a key factor. In conjunction with the aforementioned *"need"* and *"needed"*, this supports the assumption that hosts should in general provide as many amenities as possible to accommodate a larger variety of guest needs, rather than focus on catering to specific, tailored needs.

3.3 Additional Data

The dataset was also augmented with one additional data file:

1. *Nyc_subway.csv* contains information on the geographical locations of New York's subway entrances and exits, obtained from the New York state resources.

The additional data was used to model the straight distance from each listing to the nearest subway entrance, allowing us to capture subway proximity as a feature in our model if necessary. However, due to the high density of listings and subway stations, this variable was unable to sufficiently distinguish between listings. Therefore, we adjusted the variable to contain distance to the nearest major subway interchange, as well as the distance to Central Park, with their locations defined in Table 3.3.

Table 3.3 Locations

Neighbourhood	Location	*Latitude*	*Longitude*
Manhattan	Central Park	40.7829	-73.9654
Manhattan	Washington Square Station	40.7323	-74.0003
Bronx	Yankee Stadium Station	40.8282	-73.9257
Queens	Forest Hills Station	40.7216	-73.8445
Staten Island	St. George Station	40.6440	-74.0733
Brooklyn	Barclays Centre Station	40.6844	-73.9775

The distance from each listing to the nearest station and Central Park was then calculated based on the walking distance, obtained from the Google Maps API.

3.4 Occupancy Rate

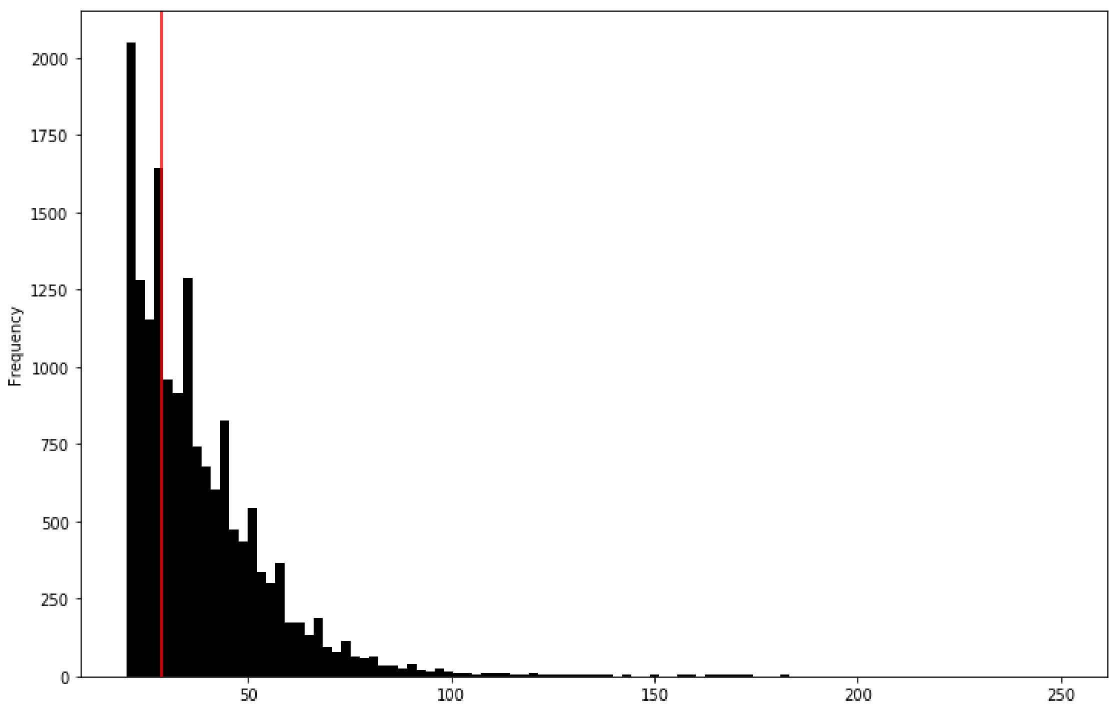


Fig3. Distribution of number of reviews per listing for number of reviews per year >= 20

Since the dataset is sourced from publicly available data, we do not have access to direct occupancy rate. Therefore, we adapted the heuristic used by Inside Airbnb to calculate occupancy rate based on the number of reviews for a listing, assuming a guest review rate of 50%, and using the average length of stay of 6.4 nights provided by Airbnb directly.

$$Occupancy\ Rate\ = (No\ of\ Reviews\ *\ 2\ *\ 6.4)/365 \qquad (1)$$

This heuristic has significant limitations. Figure 3 shows the distribution of listings with greater than 20 reviews in one year. Based on the heuristic defined, any listings with more than 28.5 reviews (shown by the red line) in a single year will have an occupancy rate of more than 100%, which is impossible. This could be a result of fake reviews used by the host to boost their ratings, or could be due to a naturally high number of short duration guests for the particularly listing. However, the identification of fake reviews does not lie within the scope of this paper, and information on average stay duration for each individual listing is unfortunately not made available to the public. Therefore, for the purpose of our research, we will simply cap the occupancy rate at 100% for listings with more than 28.5 reviews, resulting in equation (2) below.

$$Occupancy\ Rate\ = min\{No\ of\ Reviews\ *\ 2\ *\ 6.4,\ 365\}/365 \qquad (2)$$

This cap will affect 9714 listings, or 17.83% of yearly listings in our final dataset. The final breakdown of all 54476 listings is shown in Table 3.4 below.

Table 3.4 Final occupancy rate by year

	Listings	*Mean Occupancy Rate*
2016	12320	0.3567
2017	17141	0.3913
2018	25015	0.4318

4. Model Specifications

4.1 Dependent variables

In this paper, we aim to investigate significant variables affecting Airbnb business in terms of demand and supply. Occupancy rate (*occupancy_rate*) for Airbnb listings, as defined by equation 2 in section 3.4, reflects demand from customers as it measures usage of Airbnb listings during the availability period. Supply determined factor is represented in terms of listing price which is adjusted by the host in response to market demand for the listing. Natural logarithm of listing price (*log_price*) is applied to ensure the normal distribution.

Dependent Variables:
Demand-side variable: occupancy_rate
Supply-side variable: log_price

4.2 Focal variables

4.2.1 Competition

Airbnb listing faces competition from listings that have similar attributes. Listings in the same neighbourhood are often competitors of one another as they reside in the same vicinity and thus have similar environment, facilities and accessibilities the neighbourhood offers. Thus, *'house counts'* within the same neighbourhood are used to measure the level of competitiveness. When the competition is intense, it is expected the occupancy rate to be lower as there are more supplies available for consumers to choose from. Host will also likely to lower their price to attract consumers. However, it might also be argued that when the competition is intense, the areas might be a popular area of guest's choices. Thus, with higher demand, the higher the occupancy rate and price.

Focal Variables - Competition:
log_houses_count

4.2.2 Product attributes (location and amenities)

Product attributes refer to the characteristics of each listing. This can be represented by the location and amenities provided in the listings. The geographical location of the listing affects the convenience and accessibility through the travel distances to other places such as major train stations and city center. The more convenient and accessible a listing is, the more attractive it will be especially for short-term consumers who are here for touring purpose. Thus, the higher the occupancy rate. The host is also more likely to set higher price for the listing. *'dist_to_stn'* (which is the distance to the nearest major exchange train station) and *'dis_to_ctrl_park'* (which is the distance to central park, this can be perceived as the measure of distance to the city center) are used to measure the location of the listing.

Amenities show potential feature in the listings. Based on the complete list of amenities provided by Airbnb, it is grouped in the following 7 categories shown in table 4.2.2 based on the functionalities of each amenity. The more amenities a listing has, it suggests the more well-equipped the listing is, thus, increasing the attractiveness of the listing. This will potentially lead to higher occupancy rate and host setting higher prices. In our study, the well-equipped ness is measured by the percentage of amenities a listing has for each category. This is shown by the variables '*basic_amenity', 'bathroom_amenity', 'kitchen_amenity', 'children_amenity', 'convenience_amenity', 'safety_amenity', 'leisure_amenity'*. This percentage values capture the varieties of the listings. *'number_of_amenities'* is used to measure the quantity of the amenities provided.

Category	**Amenities**
Basic list	Toilet, Air conditioning,Bed linens,Essentials,Hangers,Washer,Dryer,Heating,Iron,Wifi,Hot water
Bathroom list	Bathtub,Bathtub with bath chair,Hair dryer,Shampoo
Kitchen list	Microwave,Kitchen,Oven,Stove,Refrigerator,Full kitchen,Cooking basics,Dishes and silverware,Dishwasher
Child-friendly list	Baby bath,Baby monitor,Family/kid friendly,Children's books and toys,Babysitter recommendations,Children's dinnerware,Crib,High chair,Window guards,Pack 'n Play/travel crib, Changing table
Convenience list	24-hour check-in,Building staff,Buzzer/wireless intercom,Self check-in,Elevator,Free street parking
Safety list	Lock on bedroom door,Smoke detector,Carbon monoxide detector,First aid kit
Leisure list	BBQ grill,Cable TV,Beach essentials,Breakfast,Coffee maker

Table 4.2.2 Listing categories

Focal Variables - Product Attributes:
Amenities: basic_amenity, bathroom_amenity, kitchen_amenity, children_amenity, convenience_amenity, safety_amenity, leisure_amenity, log_number_of_amenities
Location: log_dist_to_stn, log_dist_to_ctrl_park

4.2.3 Consumers' perceptions

As both an e-commerce and sharing economy platform, peer reviews on Airbnb's site nurture confidence and lubricate relations between the host and potential consumers that do not know each other. Botsman and Rogers (2010) list critical mass and trust as two important principles in the sharing economy. The use of rating systems or online verification mechanisms is a recent development, emerging from the evolution of Web 2.0 technologies to establish trust(Scott and Orlikowski, 2012). Ratings have a direct effect on the sales of goods and services (Luca and Zervas, 2016). Negative reviews on eBay (Resnick and Zeckhauser, 2002) and TripAdvisor (Scott and Orlikowski, 2012) are found to have a negative impact on price.

Thus, past guests' text reviews are used to access the impact on future occupancy choice and host's decision on price. The impact of both the quantity of the reviews and the average and variations of sentiments of the reviews are assessed by including factors *'review_when_booking'* (which measures number of reviews available when the guest is making the booking) and mean and standard deviation for *'polarity'* (which measures negativity and positivity of the review) and *'subjectivity'* (which measures objectivity and subjectivity of the review.

Focal Variables - Consumers' Perceptions:
Number of reviews: log_review_when_booking
Review sentiments: Polarity.std, Polarity.mean, Subjectivity.std, Subjectivity.mean

4.3 Control variables

To obtain robust estimates of the effects of competition, product attributes and consumers' perceptions mentioned above, we control for unobserved individual and year effects and potential confounding factors such as other product attributes, host experience and time levels.

Besides the focal product attributes variables, we also control for other significant product attributes variables, namely, the neighbourhood group in which listings are located (*neighbourhood_group_cleansed*), details of listing photos (*log_ImgBrightness, log_ImgWidth* and *log_ImgHeight*) and general product description (*log_no_words_description*). To account for quality of services provided by host based on their experience, we include two variables that measures the duration of host engaged in Airbnb business: date difference between first review and the last (*log_review_diff*) and indication of superhost (*host_is_superhost*).

Control Variables - Unobserved Variables:
Neighbourhood group: neighbourhood_group_cleansed (included as category dummies in the model)
Details of listing photos: log_ImgBrightness, log_ImgWidth and log_ImgHeight
Host experience indicator: log_review_diff, host_is_superhost

Importantly, we also include several interaction effects between focal variables to control for their simultaneous effect on the dependent variables. Interaction effect between price and house count (*log_price*log_houses_count*) is included as listing prices are set more competitively when number of listings within the same neighbourhood is high. Another interaction effect term included is between price and convenience variable (*log_price*log_dist_to_ctrl_park*) as prices are expected to be lower for listings further away from the major landmark in the city.

Control Variables - Interaction Effect:
Interaction between price and no of house in neighbourhood: log_price*log_houses_count
Details of listing photos: log_ImgBrightness, log_ImgWidth and log_ImgHeight
Host experience indicator: log_review_diff, host_is_superhost

4.4 Econometric Model Specifications

In Equation (3), we model the effects of the focal variables discussed above on listing's occupancy rate with control variables employed. The dependent variable in this model is occupancy rate of a particular listing. βs are the model coefficients of interest, α_i captures unobserved listing-specific effects, and ε_{it} is the residual error term. All the continuous variables have been log transformed in order to have a more consistent scales for the variables.

$$
\begin{aligned}
&Occupancy\_rate_{i,t} = \beta_1\, year_t + \beta_2\, neighbourhood\_group\_cleansed_i + \\
&\beta_3\, basic\_amenity_i + \beta_4\, bathroom\_amenity_i + \beta_5\, kitchen\_amenity_i + \beta_6\, children\_amenity_i + \\
&\beta_7\, convenience\_amenity_i + \beta_8\, safety\_amenity_i + \beta_9\, leisure\_amenity_i + \beta_{10}\, Polarity.std_{i,t} + \\
&\beta_{11}\, Polarity.mean_{i,t} + \beta_{12}\, Subjectivity.std_{i,t} + \beta_{13}\, Subjectivity.mean_{i,t} + \beta_{14}\, host\_is\_superhost_i + \\
&\beta_{15}\, noun\_sim_i + \beta_{16}\, adj\_sim_i + \beta_{17}\, verb\_sim_i + \beta_{18}\, log\_price_i + \beta_{19}\, log\_houses\_count_{i,t} + \\
&\beta_{20}\, log\_dist\_to\_stn_i + \beta_{21}\, log\_dist\_to\_ctrl\_park_i + \beta_{22}\, log\_number\_of\_amenities_i + \\
&\beta_{23}\, log\_review\_when\_booking_{i,t} + \beta_{24}\, log\_review\_diff_i + \beta_{25}\, log\_no\_words\_description_i + \\
&\beta_{26}\, log\_ImgBrightness_i + \beta_{27}\, log\_ImgWidth_i + \beta_{28}\, log\_ImgHeight_i + \\
&\beta_{29}\, log\_price_i \times log\_houses\_count_{i,t} + \beta_{30}\, log\_price_i \times log\_dist\_to\_ctrl\_park_i \\
&+ \theta_t + \alpha_i + \varepsilon_{it} \qquad (3)
\end{aligned}
$$

Furthermore, we are also interested to study the supply side factor price. *'Log_price'* will be the dependent variable, and *'occupancy_rate'* will substitute *'log_price'* in the independent variables for the model specification for the second study.

4.5 Measurement of Errors

R squared value is used to measure goodness-of-fit for Hausman Taylor model and random effects (RE) model as it indicates the percentage of the variance in the occupancy rate and price that the focal variables explain collectively. It used as the rough gauge of the fitness of the model.

For linear regression models, errors are measured in the form of clustered standard error in which cluster is based on listings as it accounts for correlation for the same listing across different years in the panel data. The closeness of the errors can be checked and compared among independent variables to determine their relative significances.

5. Model Estimation and Results

5.1 Identification strategy

Occupancy rate reflects consumers' demand level for a particular listings. Investigating the relationship among focal variables and occupancy rate gives insights to the effects of focal variables in determining the demand. Price, in this case, is an endogenous as it is controlled by hosts. Hosts can be perceived as suppliers in Airbnb's setting. Thus, Price can be seen as supply measure.

We used both a Hausman Taylor model and a RE specification when investigating the impact of focal variables on occupancy rate. Hausman Taylor model allows some of the regressors to be correlated with the individual effects, while RE model assumes exogeneity of all the regressors and the random individual effects (Baltagi, Bresson and Pirotte, 2002).

For price, as we only have fixed price information for each entry that do not vary across the year due to the limitation of the data, this constraints us from using fixed effects model specification. Instead, we use linear regression with clustered standard error that cluster on the dimension of individual listings effect as both as both heteroskedasticity and autocorrelation are almost certain to exist in the residuals at the individual level. Random effects specification is also employed to understand the impact of focal variables on price.

5.2 Main Analysis and results

The analysis will present the insights gathered for both demand analysis (occupancy rate modelling) and supply analysis (price modelling).

5.2.1 Occupancy Rate Modelling

The detailed results for occupancy rate modelling is shown in *table 5.2.1*. Occupancy rate has range from 0 to 1.

'houses _count' which is an indication of competition level appears to be a significant factor for occupancy rate for the FE specification. The negative coefficient indicates more intense competition is likely to lead to a drop in occupancy rate which might be due to the widely available choices when the competition in the neighbourhood is intense. The coefficient on *'log_houses_count'* is -0.015 which suggests one percent increase in houses count in the same neighbourhood will lead to 0.015% drop in occupancy rate. Furthermore, the significant interaction term between *'house_count'* and *'price'* indicates the importance of competition level in determining the ability for the host to raise a listing's price.

'dist_to_stn' which shows the location of the listing to the transportation hub appears to be a significant factor for the RE specification. The positive coefficient suggests the further away the listing is from the nearest exchange train station, the lower the occupancy rate the listing has. The coefficient on *'log_dist_to_stn'* is 0.002 which suggests one percent increase in the distance to nearest exchange train station in the same neighbourhood will lead to 0.002% increase in occupancy rate. This highlights the importance of accessibility when consumers consider whether to choose a particular listing to stay with.

For various amenities provided by the listing, child-friendly amenities, convenience amenities, safety amenities and leisure amenities stand out to be the main categories of amenities that consumers look out for when they make their choices. The coefficients are all positive except convenience amenities. Referring to *table 4.2.2*, convenience amenities include 24-hour check-in,Building staff,Buzzer/wireless intercom,Self check-in,Elevator,Free street parking. These items such as self check-in might not be well-received by all consumers as it often involves complicated process that consumers might have problems following the instructions. This potentially highlight the element of human interaction and guidance are still valued. Among all the amenities that are significant in impacting occupancy rate, child-friendly amenities appear to be the most influential factor. The coefficient on *'children_amenity'* is 0.109 for FE specification which suggests one unit increase for child-friendly amenities that a listing has will lead to 0.109 increase in occupancy rate. As *'children_amenity'* is a score ranging from 0 to 1, that means 0.1 increase for child-friendly amenities that a listing has will lead to 0.0109 increase in occupancy rate

Consumers' reviews play an important role in determining a listing's occupancy rate as all the variables related to consumers' reviews are statistically significant. *'review_when_booking'* which is a measurement of number of reviews available when the booking is made has a positive coefficient. This suggests greater number of reviews will help to establish the trust that consumers have for the host and the listing. The coefficient on *'log_review_when_booking'* is 0.212 for FE specification which suggests one percentage increase for number of reviews when the consumer made the booking will lead to 0.212% increase in occupancy rate. However, the positive coefficient for *'polarity.mean'* and negative coefficient for *'polarity.std'* are unexpected. Polarity might not be a suitable measure of reviews content. Rather, consumers might be looking out for specific content when they are reading reading the views.

Both Hausman Taylor model and RE model show similar results. The Hausman test suggests that RE are inconsistent. Thus, we prefer the Hausman Taylor model over the RE one as it allows the listing-specific unobserved heterogeneity to be correlated to the observed variable.

Table 5.2.1 Occupancy rate estimation results

	Dependent variable: occupancy_rate (range 0-1)	
	(1) FE Hausman Taylor model	(2) RE
year2017	-0.013***	-0.003
	(0.002)	(0.002)
year2018	0.026***	0.040***
	(0.002)	(0.002)

neighbourhood_group_cleansedBrooklyn	-0.013	-0.015**
	(0.013)	(0.007)
neighbourhood_group_cleansedManhattan	-0.012	-0.015**
	(0.013)	(0.007)
neighbourhood_group_cleansedQueens	-0.007	-0.009
	(0.012)	(0.007)
neighbourhood_group_cleansedStaten Island	-0.004	-0.002
	(0.020)	(0.011)
basic_amenity	0.019	0.019**
	(0.015)	(0.008)
bathroom_amenity	0.006	-0.0001
	(0.009)	(0.005)
kitchen_amenity	-0.004	-0.004
	(0.007)	(0.004)
children_amenity	0.109***	0.105***
	(0.017)	(0.010)
convenience_amenity	-0.017*	-0.017***
	(0.010)	(0.005)
safety_amenity	0.023***	0.020***
	(0.007)	(0.004)
leisure_amenity	0.020*	0.013*
	(0.012)	(0.007)
Polarity.std	0.230***	0.086***
	(0.013)	(0.013)
Polarity.mean	-0.067***	-0.040***
	(0.010)	(0.009)
Subjectivity.std	0.191***	0.072***
	(0.012)	(0.012)
Subjectivity.mean	0.118***	0.023***
	(0.010)	(0.009)
host_is_superhost	0.035***	0.027***
	(0.004)	(0.002)
noun_sim	0.232***	0.208***
	(0.025)	(0.014)
adj_sim	0.047**	0.038***
	(0.021)	(0.012)
verb_sim	0.011	0.008
	(0.011)	(0.006)
log_price	-0.054	-0.017
	(0.038)	(0.022)
log_houses_count	-0.015*	-0.003
	(0.008)	(0.005)

log_dist_to_stn	0.002	0.002^{*}
	(0.002)	(0.001)
log_dist_to_ctrl_park	-0.017	-0.006
	(0.019)	(0.011)
log_number_of_amenities	0.006	0.007
	(0.009)	(0.005)
log_review_when_booking	0.212^{***}	0.230^{***}
	(0.002)	(0.001)
log_review_diff	-0.095^{***}	-0.103^{***}
	(0.002)	(0.001)
log_no_words_description	-0.008^{***}	-0.009^{***}
	(0.003)	(0.002)
log_ImgBrightness	0.009	0.006^{*}
	(0.006)	(0.003)
log_ImgWidth	0.002	0.002
	(0.007)	(0.004)
log_ImgHeight	-0.039	-0.039
	(0.192)	(0.109)
log_price:log_houses_count	0.004^{**}	0.001
	(0.002)	(0.001)
log_price:log_dist_to_ctrl_park	0.003	0.001
	(0.004)	(0.002)
Constant	0.347	0.249
	(1.178)	(0.666)
hausmen test	p value: < 2.2e-16	chi sqr: 8114.46780242361
Observations	54,475	54,475
R^2	0.463	0.720
Adjusted R^2	0.463	0.720
F Statistic	$46,974.840^{***}$	$4,122.714^{***}$ (df = 34; 54440)
Note:	$^{*}p^{<0.1; >**}p^{<0.05; >***}p<0.01$	

5.2.2 Price Modelling

For price modelling, as prices remain unchanged over the years, linear regression model with clustered standard error and RE were applied. The detailed results for occupancy rate modelling are shown in *table 5.2.1*.

For competition variable, *'house _count'* appears to be a significant factor for price for both linear regression using clustered standard error and RE specification. The negative coefficient of -0.016 indicates that if number of listings in the neighbourhood increases by 1%, listing prices are expected to decrease by 0.016%. More intense competition is likely to lead to a decrease in listing prices which might help hosts maintain price competitiveness to attract customers.

For convenience variables, both *'dist_to_ctrl_stn'* and *'dist_to_ctrl_park'*, which show the location of the listing to the transportation hub and major landmark, appear to be significant factors for both model specifications. By interpreting coefficients, if distance between the listing and the busiest MRT exchange station increases by 1%, listing price is expected to decrease by 0.23%. If distance between the listing and major landmark increases by 1%, listing price is expected to decrease by 0.216%. This finding highlights the accessibility and location of listing are influential factors in determining listing prices.

All amenities provided seem to play an important role in determining listing prices as they are statistically significant. Coefficients for all amenities are positive except for kitchen amenity and safety amenity. Positive coefficients are reasonable as implementation of amenities incur additional cost which is reflected in the listing price. Kitchen amenity and safety amenity, on the other hand may seem mandatory which reduce bargaining power for the host. Among all the amenity types, *'children_amenity'* stands out and generates the largest positive coefficient. 1 unit increase in number of child-friendly amenities variable is expected to lead to an increase of 149.8% in listing price. This suggests children amenities are the most expensive to implement among all.

For customer review variables, *'review_when_booking'*, *'polarity.mean'* and *'subjectivity.mean'* appear to be the main factors influencing listing price. *'polarity.mean'* which is a measurement of review sentiments has the positive coefficient. Listing price is expected to increase by 18.4% if polarity score for reviews increase by 1. This suggests more positive reviews which reflect better quality of services provided will boost host confidence and set listings at a higher price. However, the negative coefficient for *'review_when_booking'* and negative coefficient for *'subjectivity.mean'* are unexpected. Number of reviews and review contents may have limited impact on listing prices as hosts may set prices in the first place before the booking. Hence, no reviews are available for their references. On the other hand, the negative coefficient for two review variables may suggest that customers give reviews more frequently for listings with higher price.

Linear regression with clustered standard error and RE model show the same coefficients and the same level of significance for all variables except for year dummy*'year_2018'*. The clustered standard errors for both models are similar which suggests robustness of the models. However, R square value suggests that competition variables, product attribute variables, and customer perception variables may have limited capacity in explaining the variation in listing prices. Other variables which reflect product values should be included to further analyse the impact on supply factors of Airbnb business.

Table 5.2.2 Occupancy rate estimation results

	Dependent variable:	
		log_price
	coefficient	*panel*
	test	*linear*
	(1)	(2)
year2017	-0.008**	-0.008
	(0.004)	(0.006)
year2018	0.028***	0.028***
	(0.005)	(0.006)
neighbourhood_group_cleansedBrooklyn	0.482***	0.482***
	(0.029)	(0.020)
neighbourhood_group_cleansedManhattan	0.598***	0.598***
	(0.028)	(0.019)
neighbourhood_group_cleansedQueens	0.363***	0.363***
	(0.028)	(0.019)
neighbourhood_group_cleansedStaten Island	0.206***	0.206***
	(0.049)	(0.031)
basic_amenity	0.355***	0.355***
	(0.033)	(0.022)
bathroom_amenity	0.211***	0.211***
	(0.021)	(0.014)
kitchen_amenity	-0.180***	-0.180***
	(0.017)	(0.011)
children_amenity	1.498***	1.498***
	(0.045)	(0.026)
convenience_amenity	0.106***	0.106***
	(0.022)	(0.015)
safety_amenity	-0.191***	-0.191***
	(0.015)	(0.010)
leisure_amenity	0.478***	0.478***
	(0.028)	(0.019)
Polarity.std	-0.040	-0.040
	(0.036)	(0.035)
Polarity.mean	0.184***	0.184***
	(0.028)	(0.026)
Subjectivity.std	-0.040	-0.040

	(0.034)	(0.032)
Subjectivity.mean	-0.063**	-0.063**
	(0.027)	(0.025)
host_is_superhost	0.035***	0.035***
	(0.008)	(0.006)
noun_sim	0.315***	0.315***
	(0.055)	(0.038)
adj_sim	-0.111**	-0.111***
	(0.044)	(0.032)
verb_sim	-0.264***	-0.264***
	(0.023)	(0.016)
occupancy_rate	-0.040***	-0.040***
	(0.013)	(0.012)
log_houses_count	-0.016***	-0.016***
	(0.003)	(0.002)
log_dist_to_stn	-0.223***	-0.223***
	(0.005)	(0.003)
log_dist_to_ctrl_park	-0.216***	-0.216***
	(0.009)	(0.006)
log_number_of_amenities	0.037	0.037**
	(0.024)	(0.015)
log_review_when_booking	-0.022***	-0.022***
	(0.004)	(0.004)
log_review_diff	0.014***	0.014***
	(0.004)	(0.003)
log_no_words_description	-0.022***	-0.022***
	(0.006)	(0.004)
log_ImgBrightness	0.137***	0.137***
	(0.013)	(0.009)
log_ImgWidth	0.161***	0.161***
	(0.014)	(0.011)
log_ImgHeight	-0.926***	-0.926***
	(0.025)	(0.296)
Constant	11.713***	11.713***
	(0.168)	(1.796)
Observations		54,475
R^2		0.334
Adjusted R^2		0.334
F Statistic		853.085*** (df = 32; 54442)

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

5.3 Robustness Check

We further corroborate our findings by checking its robustness through multiple model specifications. OLS linear regression model has been used to further validate the findings for occupancy rate model which can be referred to in *Appendix D*. The sign of the coefficients are generally consistent. While there are differences in the values for the coefficients, this is mainly due to the fact that OLS linear regression's assumption of each observation is random and independent of one another is violated in panel data.

During the modeling process, it is found that with changes in the control variables included in the model, the sign and values of coefficients are sensitive to change. Thus, appropriate and sufficient control variables are essential to ensure accurate results being obtained from the model.

We did not use another city's Airbnb data to validate the findings due to cultural and contextual differences in each city. For example, people staying in Beijing may have different preferences and neighbourhoods may be classified differently. This will subsequently affect the findings.

6. Discussion and Implications

6.1 Discussion of findings

Our study that investigates the rental prospects of Airbnb accommodation listings has several notable findings.

1. All focal variables demonstrate relative significant impact on both occupancy rate and price which represent demand and supply outcomes.
2. It is empirically shown that intense inter-market competition among Airbnb listings leads to a significant decrease on both occupancy rate and price for Airbnb listings which is shown by the negative coefficients on *'log_houses_count'*. A higher level of competition provides wider choices for customers, which leads to more elastic demand.
3. The negative coefficient for distance measures to the city center (*'log_dist_to_ctrl_park'*) and to nearest exchange train station (*'log_dist_to_stn'*) highlights the importance of listings' location. Higher accessibility and more centralized location boost both occupancy rate and price for the listings. Customer value accessibility and location which may be a driving factor for the listing prices as well.
4. The statistically significant results for most of the amenities variables suggest that most types of amenities influence both the occupancy rate and price of the business. In particular, child-friendly amenity is the most significant in boosting customers' demand as compared to other amenity as it has coefficient of the biggest magnitude. For the impact on listing prices, costs incurred for the addition of amenities are always reflected in the form of increases in listing prices which is represented by the positive coefficients for amenities variables.
5. Customers' perceptions matter differently for occupancy rate and price for Airbnb listings. Reviews exhibit a more influential role in determining occupancy rate rather than listing prices as greater number of consumer perceptions relevant variables are statistically significant in occupancy rate model. Greater number reviews improve the demand for the listings. Being the only component reflecting customer evaluation on the product, it provides informative interactions among customers to make decisions.

   However, for hosts, prices are set in prior to reviews. Time lag in information receival limits the impact of customers' perceptions on the supply side. Besides that, the difference in products valuations between customers and hosts further reduce the impact as hosts typically hold higher expectations for the value of listings than customers.

The results from our findings show similar pattern as the past studies on the effect of peer generated review content. Peer generated online review allow consumers to assess the listings more objectively, thus establishing the trust. The trust consumers place in online reviews is reflected in higher sales for businesses with better ratings, and lower sales for businesses with worse ratings (Chevalier and Mayzlin, 2003). Our study also offers insights on the quantity of reviews available in affecting both demand and supply outcomes which is not widely studied previously.

6.2 Practical Implications

Our study has several important practical implications to Airbnb markets and Airbnb hosts.

Density of Airbnb distribution should be managed to avoid overheated competition. As hosts' ability to raise price would be curtailed when there is greater competition, revenues for hosts will be compensated and in turn, decrease Airbnb's revenue. As such, Airbnb should keep in view of penetration rate for each neighbourhood and filter out less qualified listings if the rate is too high.

Evaluating product attributes allow us to understand the key factors that consumers look out for when they are making the decisions and also the factors that hosts value when they set the pricing. It is noticed that consumers and hosts generally have a similar valuation for product attributes. Locations that are central and accessible are valued highly. Thus, hosts can highlight such traits in their descriptions or titles in order to draw consumers attention. Airbnb's search results can be default prioritised by the location of the listings which are the influential factor in consumers' decision. Providing the apt choice and product that suits consumers' tastes and preference will possibly generate more successful matches.

Among amenities provided, it is worth noting that some amenities may potentially decrease the occupancy rate and price. This might be due to the mismatch in consumers' preferences. Based on the findings, child-friendly amenities and leisure amenities increase the demand for the listing and increase the bargaining power when setting prices. Thus, Airbnb can choose to send incentives for hosts that include these amenities. The form of incentives can take the form such as cash rebates or prioritising listing search results. Improving the attractiveness of the listings by equipping them with the most demanded amenities will interest more consumers to go with accommodations provided by Airbnb. Thus, generating more revenues for both Airbnb and also the hosts.

Consumers' reviews play an influential role when potential consumers make their decisions. Peer-to-Peer review content is essential in establishing the trust that consumers have for the hosts. The number of reviews a listing has is an essential evaluating factor when consumer assessing the credibility of the hosts. Thus, this can be considered to be another factor that Airbnb can employ when ranking search results, allowing listings with more reviews to be featured first. Furthermore, more incentives such as discounts for next visits or cash rebates should be given to consumers to leave a review for the listings that have visited. The accumulation of more reviews for all listings allows consumers to have more choices when trust and credibility play a key role. This thus increases the possibility for consumers to find their desired listings, increasing the revenues for both Airbnb and the hosts.

7. Conclusion

Although this research has highlighted several notable findings, we acknowledge some limitations. First, our dataset does not entail price variation in time which restrains the model choices such as fixed effects. Impact of time on prices is thus unable to capture. Apart from that, due to lack of data access, the occupancy rate was derived based on assumptions as described in section 4.1. Despite assumptions made are valid, there is still a gap between the calculated occupancy rate and the actual one. More business insights can be generated with an accurate representation of demand for Airbnb business. Moreover, our study only focus on rental prospects of Airbnb listings in New York. Impact of competition, product attributes and customers' perception variables are likely to demonstrate different impact on cities with different culture and geographic features such as Beijing.

Moving forward, we present potential avenues for future research. A meaningful extension is to evaluate the elasticity of demand and supply with respect to the three focal variables mentioned in this paper. Variables such as reviews are expected to demonstrate different values for elasticity of demand and supply. It might be also worthwhile to study customers' data. Market segmentation can be conducted to cluster customers based on their demographic data and booking behaviours. Customer study will contribute to insightful inputs for demand analysis for Airbnb as well.

8. Reference

Bivens, J. (2019, January 30). The economic costs and benefits of Airbnb: No reason for local policymakers to let Airbnb bypass tax or regulatory obligations. Retrieved April 23, 2019, from https://www.epi.org/publication/the-economic-costs-and-benefits-of-Airbnb-no-reason-for-local-policymakers-to-let-Airbnb-bypass-tax-or-regulatory-obligations/

Blal, I., Singal, M., & Templin, J. (July 2018). Airbnb's effect on hotel sales growth. *International Journal of Hospitality Management: Airbnb's Effect on Hotel Sales Growth,*85-92. Retrieved April 23, 2019.

Badi H. Baltagi, Georges Bresson & Alain Pirotte (2002). Fixed effects, random effects or Hausman-Taylor? A pretest estimator. Retrieved April 23, 2019, from http://coin.wne.uw.edu.pl/pbiernacki/ekonometria/baltagi.pdf

Hall, G. (2017, December 07). Millennial Travel Trends: Exotic on a Budget. Retrieved April 23, 2019, from https://www.huffpost.com/entry/millennial-travel-trends-exotic-on-a-budget_b_7657804

Hempel, J. (2018, July 18). Airbnb Can't Win New York-But It Can't Quit Either. Retrieved April 23, 2019, from https://www.wired.com/story/Airbnb-cant-win-new-yorkbut-it-cant-quit-either/

Haywood, J. (2016). *Airbnb and Hotel Performance: An analysis of proprietary data in 13 global markets*. Retrieved April 23, 2019, from http://www.str.com/Media/Default/Research/STR_AirbnbHotelPerformance.pdf

Juidith A. Chevalier & Dina Mayzlin. (2003). Cambridge, Mass. : National Bureau of Economic Research, 2003. The effect of word of mouth on sales.

Mody, M., & Gomez, M. (2018, October 31). Airbnb and the Hotel Industry: The Past, Present, and Future of Sales, Marketing, Branding, and Revenue Management. Retrieved April 23, 2019, from https://www.bu.edu/bhr/2018/10/31/Airbnb-and-the-hotel-industry-the-past-present-and-future-of-sales-marketing-branding-and-revenue-management/

Michael Luca and Georgios Zervas. (2016). “Fake it till you make it: Reputation, competition, and Yelp review fraud,” Management Science, volume 62, number 12, pp. 3,412–3,427. doi: https://doi.org/10.1287/mnsc.2015.2304.

Paul Resnick and Richard Zeckhauser. (2002). “Trust among strangers in internet transactions: Empirical analysis of eBay’s reputation system,” In: Michael R. Baye (editor). The economics of the Internet and e-commerce. Amsterdam: JAI, pp. 127–157.

Roach, J. C. (2018, May 01). How Airbnb has affected the hotel industry: Monthly Labor Review. Retrieved from https://www.bls.gov/opub/mlr/2018/beyond-bls/how-Airbnb-has-affected-the-hotel-industry.htm

Rachel Botsman and Roo Rogers. (2010). Whats mine is yours: The rise of collaborative consumption. New York: Harper Business.

Stewart, R. (2016, June 27). Airbnb amplifies global 'Live There' campaign with interactive cinema push and Shoreditch pop-up. Retrieved April 23, 2019, from https://www.thedrum.com/news/2016/06/27/Airbnb-amplifies-global-live-there-campaign-interactive-cinema-push-and-shoreditch

Susan V. Scott and Wanda Orlikowski. (2012). “Reconfiguring relations of accountability: Materialization of social media in the travel sector,” Accounting, Organizations and Society, volume 37, number 1, pp. 26–40. doi: https://doi.org/10.1016/j.aos.2011.11.005.

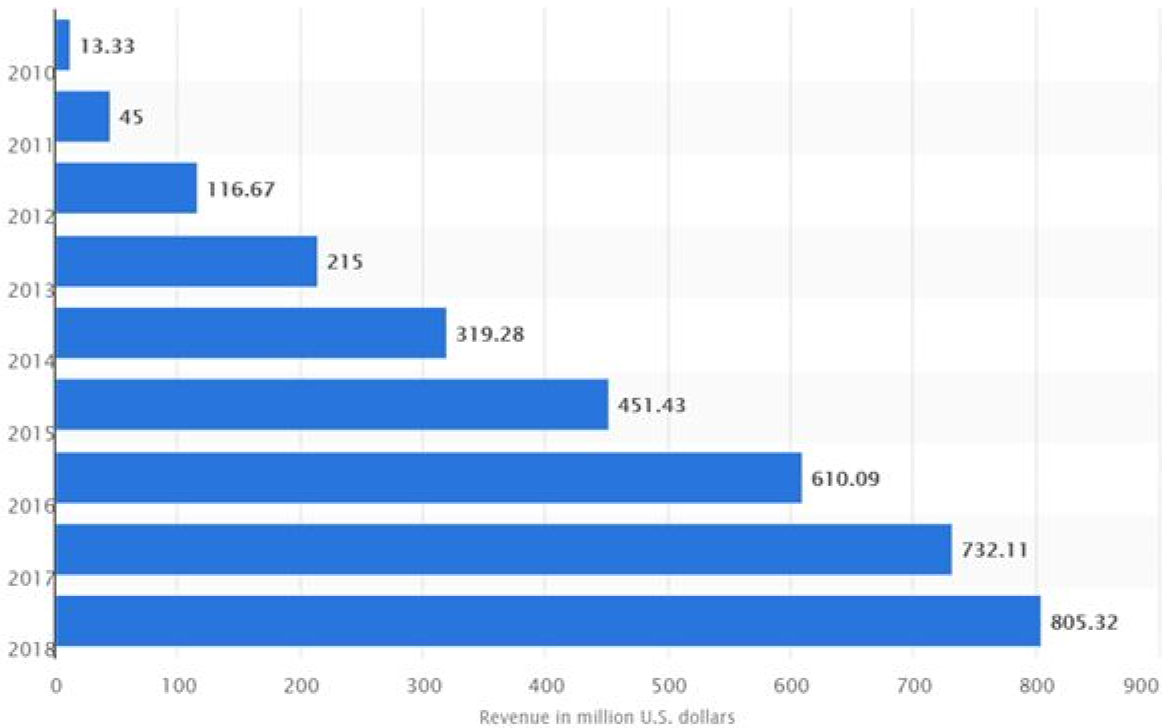
Proserpio, D. (12 april 2015). A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average. Boston University Computer Science Department.

Zervas et al, G. (22 january 2014). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Boston University School of Management*.

9. Appendix

Appendix A: Airbnb business analysis

Revenue of Airbnb in New York City from 2010 to 2018 (in million U.S dollars)
(2017 and 2018 figures are projected)



(Statista, 2016)

Value Proposition

Airbnb's value propositions are (1) to provide travelers with cheaper, unique, customized, and local experiences they can immerse themselves into when they stay in a foreign destination, and (2) to provide local owners a means to rent out their homes and generate revenues. Positioning itself to be a home away from home option, Airbnb facilitates access to local culture, activities, and people, boosting tourism especially in underexplored areas in cities. Guests gain a cheaper, more authentic travel experience, as compared to conventional hotel accommodations that are more expensive and rigid in terms of customization. Unlike mainstream hotels and hostels, Airbnb provides a plethora of accommodation choices to suit one's needs and preferences, since each accommodation listing is different from another. Outdoor glamping, treehouses, and houseboats are some unique types of accommodations Airbnb offers.

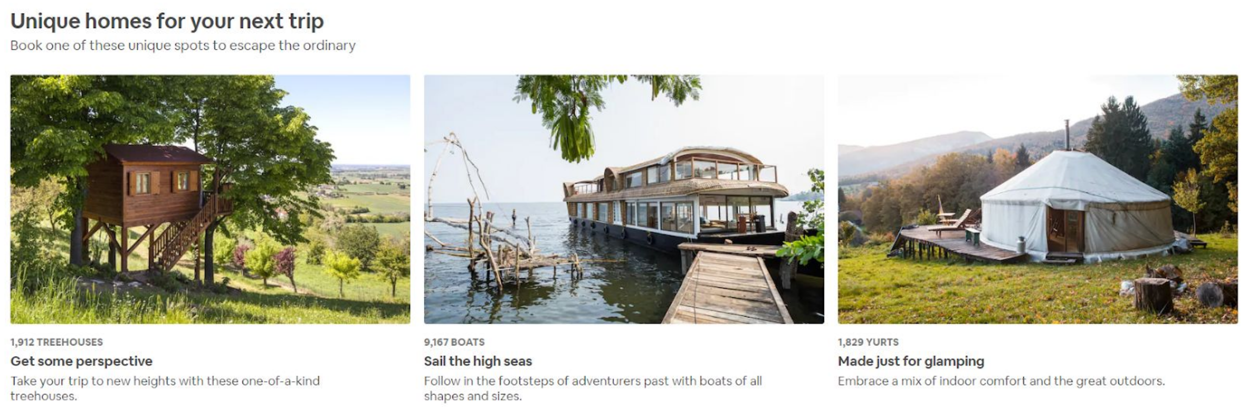


Fig2. A screenshot of listings on Airbnb's website enticing potential guests with the idea of unique unconventional homestays for an adventurous, one-of-a-kind foreign experience.

The Airbnb platform is designed in a manner that gives users the feeling that the service was created for them. The clean-design and intuitive website and mobile platforms emphasise on personalisation. They offer both counterparts ease and room for flexibility to choose their preferred requirements; guests are able to choose requirements from filters, such as price budgets, niche amenities like a seaside view, child-locks, and instant check-ins. Conversely, hosts have the liberty of pricing their apartments, and the discretion to rent on any particular day, giving them the flexibility to respond to changes in demand for accommodation. [include screenshots?] the improved mobile version has been designed to enhance user experience, to increase web traffic from mobile devices and to make booking listings more convenient for guests.

Airbnb manages to successfully mitigate the pain point of lack of trust between hosts and travelers due to information asymmetry and consumer skepticism through their peer-to-peer review model. The ability to leave reviews and ratings for hosts and guests brings about greater transparency and reduced risks between both parties.

Airbnb as a disruptive innovation

Since Airbnb's entrance into the hospitality industry, much attention has been focused on the disruptive nature of the service. Many debates and conferences have kept the spotlight on its burgeoning growth and success. It has been revealed that Airbnb's listings show substitute characteristics in their long-term effects on hotel sales' patterns. It has listed more rooms for rent than any hotel chain in the world. Airbnb's impact on the welfare gain by consumer and host is most prominent at locations and dates where hotel capacity is the most constrained, such as New

York City on New Year's Eve. Analysing the top 50 US cities with the largest hotel room inventory, research has shown that Airbnb reduced hotel profits by 3.7% in 2014, and affect hotel RevPAR (Revenue per available room) significantly. RevPAR is a performance metric used in the hotel industry, calculated by multiplying a hotel room's occupancy rate by its average daily room rate. When viewed through the lens of social welfare, this sharing economy generates $41 of consumer surplus per room per night from cost savings and $26 dollars of host surplus.

Airbnb's Business Model

Airbnb's business model is non-linear, unlike traditional hotel chains that invest in building and maintaining their properties. Airbnb does not own any properties, and gets its share of revenues from both of its customers: Airbnb receives a flat 10% commission from hosts for each confirmed booking, and a 3% payment processing fee. It receives 6-12% of the total booking amount from guests, set as a non-refundable service fee. Both hosts and guests are incentivized by Airbnb's rating system to use signaling mechanisms to build trust and maximize the likelihood of successful bookings .

Strengths	Weaknesses
1) relatively low costs due to the absence of ownership of any accommodations 2) Excellent customer-centric service for both hosts and guests 24/7 3) dual rating system facilitates trust, safety, and transparency between hosts and guests 4) successful influencer marketing: people are much more likely to accept a brand if a famous person has done so, such as Lady Gaga and Mariah Carey endorsing Airbnb organically (Jones, 2017). E.g. in 2016, Airbnb launched a "Don't go there, live there" London campaign, enlisting 25 influencers to show travelers the "real London" not shown in travel guidebooks (Stewart, 2016).	1) heavy reliance on quality hosts deliver and satisfaction of host experience → lack control over its internal performance 2) Airbnb incurs additional costs from disputes between hosts and guests

Opportunities	Threats
1) Dissatisfaction with hotels due to high prices and less than satisfactory accommodations conditions leads to an increase in demand in Airbnb accommodations (substitute) 2) Increase in take up rate of mobile devices: Airbnb taps on the increasing reliance on mobile devices to provide a convenient and personalised method of booking accommodations 3) Increasing popularity of alternative or budget traveling among the younger generation (Hall, 2017)	1) Demand for listings vulnerable to economic turbulence 2) Increasing number of competitors due to easily replicated concept e.g TripAdvisor, HomeStay, Flipkey 3) Hotels and hostels evolving to compete for Airbnb's traveler consumer base by providing customized accommodation experiences (Mody, 2018 4) Rising cases of lobbying against Airbnb and tax regulations: NewYork City's Multiple Dwelling Law (Hempel, 2018) → affects Airbnb's reputation and image negatively.

Appendix B: Part-of-Speech Tag Descriptions

Tags	**Description**	**Examples**
VB	Verb, base form	*Take*
VBD	Verb, past tense	*Took*
VBG	Verb, present participle	*Taking*
VBN	Verb, past participle	*Taken*
VBP	Verb, singular present	*Take*
VBZ	Verb, third person singular present	*Takes*
NN	Noun, singular	*Table*
NNS	Noun, plural	*Tables*
NNPS	Proper noun, singular	*John, America, etc.*
NNP	Proper noun, plural	*Americans*
JJ	Adjective	*Big*
JJP	Adjective, comparative	*Bigger*
JJS	Adjective, superlative	*Biggest*

Appendix C: Data description for model input

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
year	54,476	2,017.2	0.8	2,016	2,017	2,018	2,018
listing_id	54,476	12,465,764.0	8,351,323.0	2,539	4,936,254	19,246,674.0	31,201,624
price	54,476	140.6	161.0	0	70	170	10,000
occupancy_rate	54,476	0.4	0.4	0.04	0.1	0.7	1.0
houses_count	54,476	656.2	595.1	1	178	959	2,106
dist_to_stn	54,476	5,165.7	3,363.8	4	2,427.8	7,030	25,446
dist_to_ctrl_park	54,475	9,901.6	5,418.2	576.0	5,570.0	13,649.0	44,040.0
number_of_amenities	54,476	23.0	9.9	1	16	29	82
basic_amenity	54,476	0.6	0.2	0.0	0.5	0.7	0.9
bathroom_amenity	54,476	0.4	0.2	0.0	0.2	0.5	1.0
kitchen_amenity	54,476	0.3	0.3	0.0	0.1	0.7	1.0
children_amenity	54,476	0.1	0.1	0	0	0.1	1
convenience_amenity	54,476	0.2	0.2	0.0	0.2	0.3	1.0
safety_amenity	54,476	0.6	0.3	0.0	0.5	0.8	1.0
leisure_amenity	54,476	0.1	0.2	0.0	0.0	0.2	1.0
Polarity.std	54,476	0.2	0.1	0.0	0.1	0.2	1.2
Polarity.mean	54,476	0.4	0.1	-1.0	0.3	0.4	1.0
Subjectivity.std	54,476	0.2	0.1	0.0	0.1	0.3	0.7
Subjectivity.mean	54,476	0.6	0.1	0.0	0.5	0.6	1.0
review_when_booking	54,476	43.6	57.6	0	5	61	587
no_months_host	54,476	58.3	22.4	14	41	75	126
review_diff	54,476	27.5	22.1	0	10	40	119
host_is_superhost	54,476	0.3	0.4	0	0	1	1
no_words_description	54,476	146.2	46.5	0	135	176	219
ImgBrightness	54,476	130.2	30.9	6.0	107.6	152.4	238.4
ImgWidth	54,476	593.0	104.4	205	639	639	2,560
ImgHeight	54,476	426.0	6.5	426	426	426	1,503
noun_sim	54,476	0.1	0.1	0.0	0.1	0.2	0.5
adj_sim	54,476	0.1	0.1	0	0	0.1	1
verb_sim	54,476	0.2	0.1	0	0.1	0.3	1
log_price	54,476	4.7	0.6	0.0	4.2	5.1	9.2
log_houses_count	54,476	5.9	1.3	0.0	5.2	6.9	7.7
log_dist_to_stn	54,476	8.3	0.8	1.4	7.8	8.9	10.1
log_dist_to_ctrl_park	54,475	9.0	0.7	6.4	8.6	9.5	10.7
log_number_of_amenities	54,476	3.0	0.5	0.0	2.8	3.4	4.4
log_review_when_booking	54,476	2.8	1.6	0.0	1.6	4.1	6.4

log_no_months_host	54,476	4.0	0.4	2.6	3.7	4.3	4.8
log_review_diff	54,476	2.8	1.3	0.0	2.3	3.7	4.8
log_no_words_description	54,476	4.9	0.6	0.0	4.9	5.2	5.4
log_ImgBrightness	54,476	4.8	0.3	1.8	4.7	5.0	5.5
log_ImgWidth	54,476	6.4	0.2	5.3	6.5	6.5	7.8
log_ImgHeight	54,476	6.1	0.01	6.1	6.1	6.1	7.3

Appendix D: Robustness check for occupancy rate model

	Dependent variable: occupancy_rate		
	OLS	*panel linear*	
	(1)	(2)	(3)
year2017	-0.003	-0.013***	-0.003
	(0.002)	(0.002)	(0.002)
year2018	0.040***	0.026***	0.040***
	(0.002)	(0.002)	(0.002)
neighbourhood_group_cleansedBrooklyn	-0.015**	-0.013	-0.015**
	(0.007)	(0.013)	(0.007)
neighbourhood_group_cleansedManhattan	-0.015**	-0.012	-0.015**
	(0.007)	(0.013)	(0.007)
neighbourhood_group_cleansedQueens	-0.009	-0.007	-0.009
	(0.007)	(0.012)	(0.007)
neighbourhood_group_cleansedStaten Island	-0.002	-0.004	-0.002
	(0.011)	(0.020)	(0.011)
basic_amenity	0.019**	0.019	0.019**
	(0.008)	(0.015)	(0.008)
bathroom_amenity	-0.0001	0.006	-0.0001
	(0.005)	(0.009)	(0.005)
kitchen_amenity	-0.004	-0.004	-0.004
	(0.004)	(0.007)	(0.004)
children_amenity	0.105***	0.109***	0.105***
	(0.010)	(0.017)	(0.010)
convenience_amenity	-0.017***	-0.017*	-0.017***
	(0.005)	(0.010)	(0.005)
safety_amenity	0.020***	0.023***	0.020***
	(0.004)	(0.007)	(0.004)
leisure_amenity	0.013*	0.020*	0.013*
	(0.007)	(0.012)	(0.007)
Polarity.std	0.086***	0.230***	0.086***
	(0.013)	(0.013)	(0.013)
Polarity.mean	-0.040***	-0.067***	-0.040***
	(0.009)	(0.010)	(0.009)
Subjectivity.std	0.072***	0.191***	0.072***
	(0.012)	(0.012)	(0.012)
Subjectivity.mean	0.023***	0.118***	0.023***
	(0.009)	(0.010)	(0.009)
host_is_superhost	0.027***	0.035***	0.027***

	(0.002)	(0.004)	(0.002)
noun_sim	0.208***	0.232***	0.208***
	(0.014)	(0.025)	(0.014)
adj_sim	0.038***	0.047**	0.038***
	(0.012)	(0.021)	(0.012)
verb_sim	0.008	0.011	0.008
	(0.006)	(0.011)	(0.006)
log_price	-0.017	-0.054	-0.017
	(0.022)	(0.038)	(0.022)
log_houses_count	-0.003	-0.015*	-0.003
	(0.005)	(0.008)	(0.005)
log_dist_to_stn	0.002*	0.002	0.002*
	(0.001)	(0.002)	(0.001)
log_dist_to_ctrl_park	-0.006	-0.017	-0.006
	(0.011)	(0.019)	(0.011)
log_number_of_amenities	0.007	0.006	0.007
	(0.005)	(0.009)	(0.005)
log_review_when_booking	0.230***	0.212***	0.230***
	(0.001)	(0.002)	(0.001)
log_review_diff	-0.103***	-0.095***	-0.103***
	(0.001)	(0.002)	(0.001)
log_no_words_description	-0.009***	-0.008***	-0.009***
	(0.002)	(0.003)	(0.002)
log_ImgBrightness	0.006*	0.009	0.006*
	(0.003)	(0.006)	(0.003)
log_ImgWidth	0.002	0.002	0.002
	(0.004)	(0.007)	(0.004)
log_ImgHeight	-0.039	-0.039	-0.039
	(0.109)	(0.192)	(0.109)
log_price:log_houses_count	0.001	0.004**	0.001
	(0.001)	(0.002)	(0.001)
log_price:log_dist_to_ctrl_park	0.001	0.003	0.001
	(0.002)	(0.004)	(0.002)
Constant	0.249	0.347	0.249
	(0.666)	(1.178)	(0.666)
Observations	54,475	54,475	54,475
R^2	0.720	0.463	0.720
Adjusted R^2	0.720	0.463	0.720
Residual Std. Error	0.194 (df = 54440)		
F Statistic (df = 34; 54440)	4,122.714***	46,974.840***	4,122.714***
Note:	*p$^{<0.1; >**}$p$^{<0.05; >***}$p<0.01		