Impact of Price on AirBnb Bookings in Los Angeles

What can the Los Angeles Tourism & Convention Board learn about Airbnb booking behavior?

Austin Jin, Lynn Marciano, Jacquie Nesbitt, Huyette Spring

Contents

1 Introduction

Throughout the year 2020, tourism and hospitality industries throughout the world have been tremendously impacted by the COVID-19 pandemic. Government bodies, healthcare workers, and even celebrities urged people in Los Angeles to "stay home and stay strong" in order to prevent the spread of the contagious virus. With that being said, there have been confirmed reductions in infections which led to airlines reopening on a massive scale. "Revenge travel", a term used for people who are now more eager to travel after a year of putting trips on hold, has hence become a frequently used phrase in our vocabularies. By understanding the factors that lead people to make booking decisions, our team would be able to provide better guidance to current or potential Airbnb hosts to help the Los Angeles's tourism and hospitality industries capitalize on higher interest.

To further prepare for the upward trend in bookings, our research study will focus on finding the components that contribute to booking 60 days out through Airbnb. More specifically, our team has decided to analyze the relationship between Airbnb prices and the number of bookings within Los Angeles County to answer the following research question:

How does price affect Airbnb bookings?

Through our analysis, we will be establishing whether Airbnb prices impact bookings to provide insights into the determinants that travelers use to finalize booking decisions. Armed with the insight we gained into booking behavior, we aim to set the foundation for future policy interventions and research which the Los Angeles Tourism & Convention Board might leverage in order to help hosts increase their bookings and ultimately increase tourism in the city.

2 Model Building Process

2.1 Data Description

To operationalize our research question, we utilized data from Inside Airbnb. Inside Airbnb scrapes publicly available information about a city or county's Airbnb listings. The accuracy of the information compiled from the Airbnb site is not controlled by Inside Airbnb. However, Inside Airbnb has analyzed, cleansed, and aggregated the compiled data where appropriate.

It is important to note that Inside Airbnb is an advocacy group that advocates for reducing Airbnb short-term rentals. Although "due care" has been taken with the processing and analysis of the dataset, we remain cautious about the cleanliness and objectivity of this data, the advocacy group's motivation for collecting this data, and the risk of listings being removed after data cleansing.

2.2 Exploratory Data Analysis

Based on our step-by-step assumptions on the key motive variables that would lead individuals to book a listing on Airbnb, we have decided to utilize the following variables for our analysis:

- availability_60
- price
- room_type
- reviews_scores_rating
- accommodates
- instant_bookable

Filtering

When assessing our variables, we wanted to define who a vacationer would be. Typically, individuals will book for a short period of time when vacationing. However, we found listings where the minimum number of nights for a booking was greater than 60 days. This led us to assume that these listings may be used more so for rental purposes rather than vacationing. We decided to filter our data to remove listings with more

than a 60 day minimum booking. It is also worth mentioning that travelers usually traveled for 2 weeks or more but now with COVID-19 reshaping the travel industry, travelers are working remote and staying in places longer than usual. Thus, we have decided to remove all listings that did not fit this criteria.

Furthermore, when looking at the "availability_60" variable and "has_availability" variable, we noticed a discrepancy between the two. The "has_availability" variable states where there is availability for booking the listing. However, we found that the "availability_60" and the other "availability_n" would be 0, which states there is no availability, and the listing is fully booked. We decided to remove the rows where we found this discrepancy.

availability_60

The "availability_60" variable represents the availability of the listing 60 days into the future as determined by the calendar from the data scrape date. Although there were other "availability_x" variables, we decided to focus on two months out from the calendar scrape date. We recognize that peak tourism season is between June and August which may impact price but we consider the unselected timeframe as an omitted variable. Understanding the outcome variable of bookings will allow us to provide insight to potential and current Airbnb hosts to increase their bookings. This outcome variable allows us to operationalize the concept of demand in bookings 60-days in advance. It is important to note that a listing may not be available due to being booked by a guest or blocked by the host. The "availability_60" variable will be used to analyze how many bookings are truly booked 60 days out into the future.

price

The "price" variable represents the current nightly price for the listing at the time of the scrape date, assuming that the price will be the same for all the days in the calendar dates. The variable also does not include any additional charges that are included when booking a listing, such as: cleaning fees, booking fees, pet fees, and any discounts. We wanted to focus our research on price because we believe that price is the most important motive when individuals book a listing on Airbnb.

When assessing the "price" variable, we observed that it is right-skewed. This confirms our interpretations that listings are typically moderately priced and only few are abnormally priced higher than the average. Applying a base10 log transform to the variable makes it normally distributed for our statistical analysis.

room_type

The "room_type" variable represents the type of listing grouped into the following categories:

- Entire place
- Private room
- Hotel room
- Shared room

There are different reasonings that would motivate an individual to book a specific type of listing. Based on the unknown subjective nature of individuals, all four categories were kept as factored variables. Entire Places may be more sought after due to having the whole space to yourself. With Private Rooms, an individual still has some of their own space but may share corridors with others, such as the host, or other travelers. A Shared Room is where the individual will be sleeping in a shared space with others. Based on our hypothesis of social awareness from the pandemic, we recognized why this variable is not highly represented in our dataset. Hotel Rooms offer a private space similar to that of an Airbnb private room but with the convenience of online booking for most hotels, it may be more likely that an individual would book on platforms other than Airbnb. Entire Places carried the most representation in this variable, followed by Private Room, Shared Room, and Hotel Room.

reviews_scores_rating

The "reviews_scores_rating" variable represents the ratings score for listings from 0 (worst) to 5 (best) for the overall experience and for specific categories, including: overall experience, cleanliness, accuracy, value, communication, check-in, and location. We felt this variable was important as a motive behind booking on Airbnb due to the assumption that individuals would want a quality stay when traveling.

When assessing the variable, we observed that it is left-skewed. It was in fact our heaviest skewed variable. Transforms such as log and poly-normalizing the variable showed no effect, so we left the data distribution as-is to best represent the data. However, a rating from 1-2 is not the same as 4-5 due to the subjective nature of individuals based on their experience during their stay.

When assessing the relationship between price, rating, and room type. We were able to observe that private stays, such as, "Entire homes/apartments", "Hotel Rooms", and "Private Rooms" were indeed priced higher than "Shared Rooms". Many of the listings are clustered around higher ratings; however, there is a linear declination in price when the ratings are higher. This led us to assume that listings with better ratings could be a result of being valued at a better price. Therefore, we decided to explore the effects of these variables in conjunction in our models.

accommodates

The "accommodates" variable represents the maximum capacity of the listing. In order words, the variable provided us with data on the number of guests that could be housed for the given listing. We recognized this variable as a criterion when booking an Airbnb reservation and intended on seeking whether there was any relationship between accommodations coupled with price and room type with the availability of the listing 60 days out. When observing the distribution of the capacity of Airbnb bookings, we notice that it was right-skewed. This tells us that most listings are capable of hosting 8 or fewer guests and there are very few listings that are able to accommodate more.

instant_bookable

The "instant_bookable" variable represents whether the guest can automatically book the listing without the host being required to accept their booking request. This variable is considered a potential indicator of a commercial listing. It is also worth mentioning that "instant_bookable" is a boolean data type in that "t" would equal true for the guest being able to automatically book the listing without host approvals and "f" for the guest not being able to automatically book the listing without host approvals.

We felt this variable would be important to include in our models due to the value of convenience in humans' daily lives. We wanted to see if the convenience factor of being able to instantly book a listing without discussing with a host would increase the likelihood of a listing being booked. When assessing the listings that were instant_bookable, we observed that more listings were not instant_bookable. This led to our assumption that most listings on Airbnb are privately owned, rather than commercially owned.

2.3 Model Selection and the Subset Selections

Our models were structured under the notion of assumptions that would motivate individuals to book certain Airbnbs. The general framework considered thinking patterns that would drive motivations for individuals to book on Airbnb. Individual motivations also included the built-in features that the Airbnb platform has to offer. Below is a brief explanation for all 4 models:

Model 1 Model 1 is our limited model that assesses the relationship between our key variable "price" and "availability_60". In essence, we intended to measure how price affects bookings that are 60 days out into the future. Before applying any transformations, price was right-skewed. We applied a base10 log transformation to the variable and the distribution became normal.

Model 2 Model 2 includes some additional explanatory variables that we believed were other motives for booking through Airbnb. We assumed that the type of listing ("room_type") and rating ("reivews_scores_rating") were part of the primary framework when an individual performed bookings on Airbnb.

Model 3 Model 3 includes another covariate that we believed would be part of the booking process. Knowing that the size of a booking party is considered an important criterion when filtering results through Airbnb, we have decided to add in the "accommodates" variable for specific bookings that are based on the accommodation capacity of a listing.

Model 4 Model 4 includes another covariate that we believed would drive more bookings. Knowing that individuals prefer convenience when it comes to being able to seamlessly book through an online platform, we have decided to also include the "instant_bookable" variable as a factor that would affect the number of bookings.

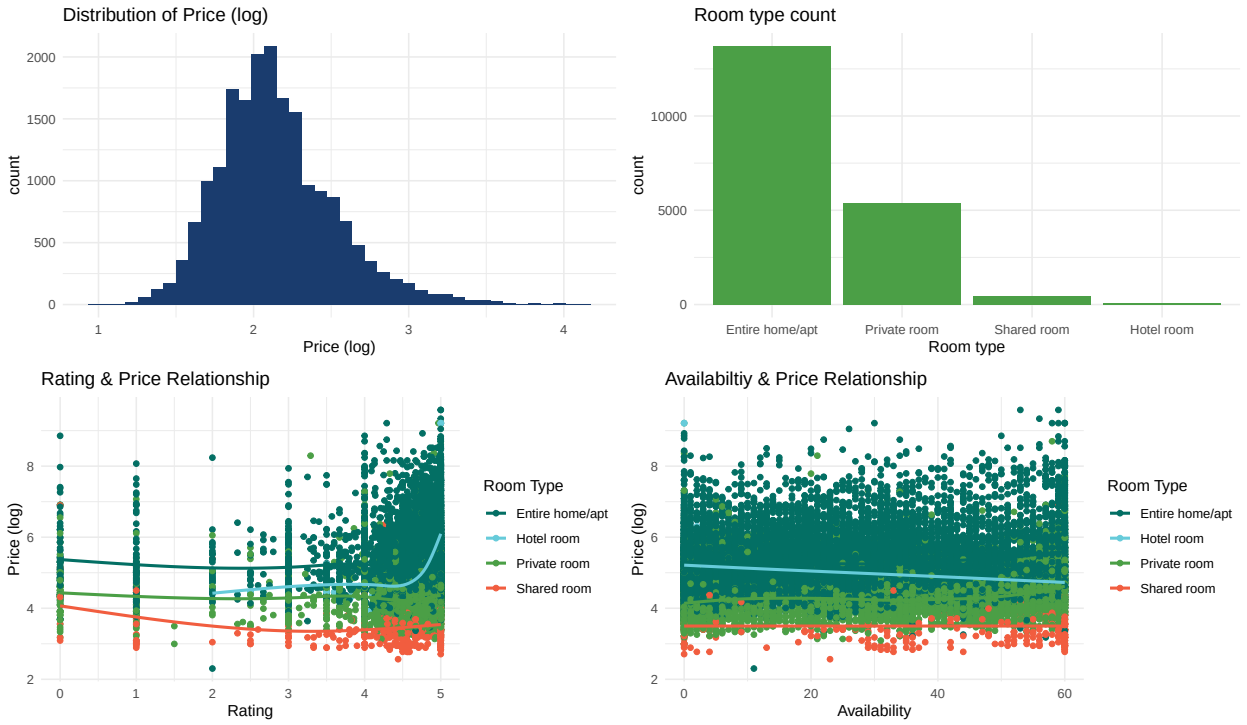

Figure 1: AirBnb Bookings Exploratory Data Analysis

3 Regression Results

Regression Results with Standard Robust Errors

	Dependent variable:			
	Availability 60 days In The Future			
	Price Only	+ Room Type and Rating	+ Accommodates	+ Instantly Bookable
	(1)	(2)	(3)	(4)
Log of Price	2.7***	11.7***	14.6***	14.7***
	(0.4)	(0.5)	(0.6)	(0.6)
Hotel Room		19.5***	18.6***	17.3***
		(2.8)	(2.9)	(2.9)
Private Room		10.1***	9.7***	9.8***
		(0.4)	(0.4)	(0.4)
Shared Room		31.0***	31.8***	31.5***
		(1.0)	(1.0)	(1.0)
Overall Rating		-4.2***	-4.2***	-4.1***
		(0.3)	(0.3)	(0.3)
Number of People the Listing Accommodates			-0.7***	-0.7***
			(0.1)	(0.1)
Instantly Bookable				2.0***
				(0.3)
Constant	20.2***	16.6***	13.3***	12.0***
	(0.9)	(1.7)	(1.7)	(1.7)
Observations	19,647	19,647	19,647	19,647
R2	0.002	0.1	0.1	0.1
Adjusted R2	0.002	0.1	0.1	0.1
Residual Std. Error	20.5 (df = 19645)	19.5 (df = 19641)	19.5 (df = 19640)	19.4 (df = 19639)
F Statistic	47.7*** (df = 1; 19645)	409.3*** (df = 5; 19641)	354.6*** (df = 6; 19640)	311.3*** (df = 7; 19639)
Note:				*p<0.1; **p<0.05; ***p<0.01

Based on the large sample assumptions, we are only required to fulfill the I.I.D and a unique BLP exists. There are reasonably heavy tails in the data. Where applicable, we applied transformations to our variables, specifically our primary variable, "price", which made the variable normally distributed. However, applying transformations to our supporting variables showed no effect. The log transformation did reduce the impact of the heavy tails on our primary variable and no perfect collinearity (as discussed in our CLM assumptions), therefore we assume that a unique BLP does exist.

According to the regression results, one thing that stands out is that all variables are statistically significant and the coefficients are somewhat stable across model 2, model 3, and model 4. Overall, the main takeaway from these regressions is that all models do find a statistically significant relationship between price and a listing's availability 60 days into the future. In fact, price was materially impactful to bookings; a 1% increase in price leads to a more than 14 days decrease in bookings while holding all else constant. Given the variables included in model 4, the size of the coefficients associated with price demonstrates that like-for-like Airbnb bookings are a competitive market.

However, we urge caution when analyzing these statistics since the outcome variable used also includes days the host may have made unavailable and not the result of a customer.

Additionally, we learned the following from the other variables that may impact our variable of interest:

- 16 days available resulting from the listing being a whole apartment/house when all other variables are set to 0.
- 17 days available resulting from the listing being a hotel room
- 9 days available resulting from the listing being a private room
- 31 days available resulting from the listing being a shared room
- .7 days booked for every person the listing says it accommodates
- 4 days booked for every unit increase of the overall rating

- 2 days available for having the instantly bookable feature

However, given the low adjusted R-squared, we believe these are not the only variables that impact a listing's availability 60 days into the future since they do not fully capture the variance in our dataset.

Across models 2, 3, and 4, we believe Model 4 is the one that would best capture potential effects because of its higher adjusted R-squared as compared to the other model specifications. Additionally, we ran ANOVA which determined that each model was statistically significant compared to its predecessor. That being said, due to the recursive nature of the booking behavior, we considered the more complex model which includes interaction terms. While this model boosted the adjusted R-squared and the variable of interest was less by 1 day than the non-interaction model, we did not feel the additional complexity was worth it, especially in light of the low adjusted R-squared, its interpretation complexity, and it didn't help meet the CLM assumptions.

4 CLM Assumptions and Limitations of the Model

Given the number of observations in the dataset (n=19,647), we believe it is reasonable to rely on the Central Limit Theorem. However, for the purposes of articulating the limitations of our model, we evaluate the five Classic Linear Model assumptions. As demonstrated below, our model fails a number of these tests, providing justification for the use of Robust Standard Errors.

4.1 IID Sampling

There are several characteristics of the data to note when evaluating IID:

Inside Airbnb is an advocacy group. While Inside Airbnb may not have altered the data, there is no way to validate this. At a minimum, we relied on this advocacy to provide the cleaned data from the scraping process. Accordingly, we must assume that the data has some degree of bias, violating the assumption of IID, albeit to an unknown degree.

Supply / Demand Recursion. Airbnb is a marketplace, balancing supply and demand through price. In this marketplace, landlords compete to have their rooms booked. Considering market conditions, the price and availability of one room will impact the price and availability of another, in a recursive loop characteristic of all markets. Therefore, while the data only represents a point in time (as opposed to a time-based study), the nature of the data violates a core assumption of IID sampling, namely, independence.

Geography. As representatives of the Los Angeles Tourism & Convention Board, we are focused on the City of Los Angeles, specifically. While this is a reasonable and practical approach for the study, it does not represent random sampling. The renter and rentee behaviors within Los Angeles, or even within neighborhoods in Los Angeles, are unlikely to be truly independent from one another. While potentially unavoidable, this does limit the assumption of IID.

Based on the above, we make a specific note of the limitations of the IID assumption for this dataset. However, we also note that it may be nearly impossible, and at a minimum impractical, to remove these limitations. Accordingly, for the purposes of this study, we proceed under the assumption that the data is IID.

4.2 No Perfect Colinearity

A simple indication of perfect collinearity is that regressions won't run, or will drop a variable. Based on the Regression Results noted above, we can see that the regression ran appropriately and that no values were dropped (displayed as N/A). Additionally, running a VIF test did not reveal any scaling values above 5. Accordingly, we can assume that the model does not include any perfectly collinear variables.

4.3 Linear Conditional Expectation

When assessing the Linear Conditional Expectation, we aim to evaluate whether the data we have can be described in terms of a linear model. This assumption is important because it's one of three assumptions

that tell us if we can trust the coefficients of the model. From the "ocular test", in which a regression is fit on the scatter plot of prediction against residuals, demonstrates a distinct declining slope in which higher predictions have a significant negative residuals bias. Moreover, the residuals associated with price and review score also notably deviate from zero. This indicates that a model inclusive of polynomial transformations may be more appropriate but is beyond the scope of this study.

Based on the above, we do not believe we have satisfied the assumption of Linear Conditional Expectation.

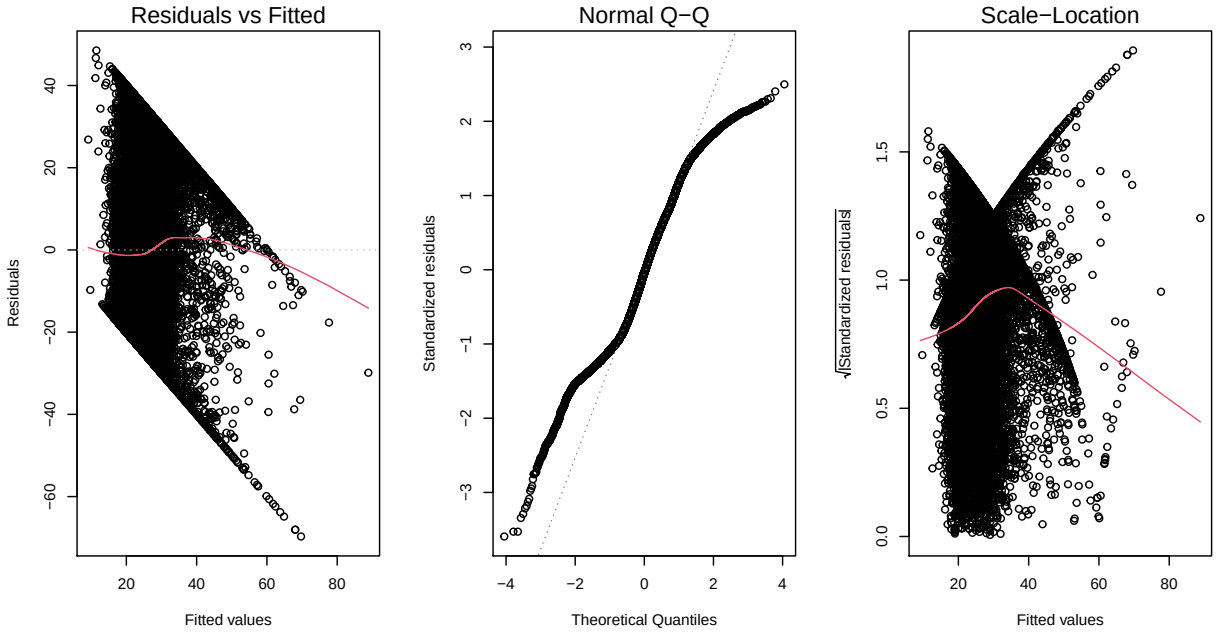

Figure 2: Model 4 CLM Assumptions

4.4 Homoskedastic Errors

Homoskedastic Errors refers to constant, finite conditional error variances. If the conditional error variances differ (heteroskedastic), it could indicate, mainly, that Classical Standard Errors are biased. Heteroskedastic variance is often caused by, among other things, skewness in the data.

In the case of the Airbnb data, we have already noted the skewness of the variables and therefore it is easy to imagine that they have Heteroskedastic variance.

There are two approaches to assessing Homoskedastic Errors: the Breusch-Pagan test, and the "ocular" test. The scale location plot demonstrates a distinct curve indicating that the data contains heteroskedastic errors.

The ocular test is confirmed by the Breusch-Pagan test which has a materially significant p-value of 2.2e-16, suggesting that we reject the null hypothesis of homoskedasticity in favor of heteroskedasticity.

Based on these tests, we do not believe we have satisfied the assumption of Homoskedastic Errors. And therefore, we have used robust standard errors in our regression analysis.

4.5 Normally Distributed Errors

When evaluating Normally Distributed Errors, we are attempting to demonstrate that the residuals of the model are normally distributed around 0. Normally distributed errors are important because we need them to trust the hypothesis testing results. The Normal Q-Q chart demonstrates material deviations in the

Theoretical Quantiles away from the straight-line fit which would indicate a normal distribution. It is worth mentioning that there are tails instead of a straight diagonal line. .

Again, we believe that the assumption of Normally Distributed Errors has not been met.

5 Ommitted Variables

The main relationship we explored in our model was between price and a listing's availability 60 days into the future. As noted above, Airbnb is a marketplace, balancing supply and demand through price in a recursive loop characteristic of all markets. Because of this, identification of omitted variables that are not direct outcomes of this recursive balancing of supply and demand is a challenge. In other words, the omitted variables identified must impact availability and price simultaneously rather than influencing one which subsequently influences the other. Nevertheless, we have attempted to identify such omitted variables below:

COVID-19. In terms of causality, we could easily argue that COVID-19 could cause a decrease in bookings as people stay home. Likewise, we could also argue that it would cause an increase in bookings because people are working remotely and are taking the opportunity to live away from their primary residence longer. We believe this to be a bi-directional variable that could influence the direction of the bias towards or away from zero.

Environmental conditions. We could argue that due to the nearby wildfires happening near Los Angeles the air quality would give potential tourists pause for booking a potential holiday in the area. If there's a positive relationship between our 60 days availability and wildfires and a negative relationship between price and wildfires, then the true coefficient of price would be towards zero.

General economic conditions. The general economic conditions of both Los Angeles and the home location of the Airbnb booker could be an omitted variable, biasing availability and price. Positive economic conditions may decrease availability as bookers have more disposable income while also increasing price as landlords anticipate this additional disposable income. Note that, as mentioned above, if landlords increased price as a result of and subsequent to a reduction in supply, this would not be an omitted variable but rather a function of supply and demand balancing through price.

Seasons (low and shoulder). Since the price variable was scrapped for July 6th and 7th, it only accounts for the peak tourist season. If there's a positive relationship between the number of days available and the lower and shoulder seasons and a negative relationship between the low and shoulder seasons and price then the true coefficient of price would be towards zero.

6 Conclusion

Background and goals. The group set out to identify drivers that current and potential Airbnb hosts may be able to leverage to capture more revenue from "revenge traveling" tourists.

Findings. We hypothesized that there is a relationship between price and availability within 60 days. At a high level, our model validated that there is a positive relationship between price and availability within 60 days. Additionally, we found that price, room type, ratings, accommodates, and instant bookability were all statistically significant.

Next steps. We encourage the Los Angeles Tourism & Convention Board to support future research such that we can identify concrete actions to positively impact these variables. This future research may include:

(i) surveys and in-person interviews of both Airbnb renters and rentees to gain a better understanding of actual behavior;

(ii) an enhanced model which captures the findings from these surveys and interviews, potentially including interaction variables; and

(iii) experimental trials to determine the effectiveness of recommended interventions.

7 Bibliography

- Data Source: Inside Airbnb Listings Data
- Data Dictionary: Inside Airbnb Data Dictionary
- Forbes - Revenge Travel And Where Americans Are Traveling
- Fortune - Airbnb CEO Brian Chesky: Expect longer trips after COVID-19