ANALYSIS OF AIRBNB DATA FROM RIO DE JANEIRO (APRIL 2018)

APRIL 29, 2022

[redacted]

Non-Technical Summary

Airbnb operates by charging guests and hosts for short-term rental stays in private accommodation that are booked through the Airbnb website. Airbnb has grown to 5.6 million active listings and 4 million hosts[1]. It's usually depicted as a good option for travellers looking for low-cost accommodations, as well as for homeowners looking for extra income. Guests are asked to write a review within 14 days of their stay.

The type of room of Airbnb was found to have a big impact on price, with the cost of an entire home being higher than the cost of a shared room. Further, hosts that consider their property as providing better amenities, such as the number of bathrooms and accommodation capacity, will charge higher prices. A higher number of guests included in the price has a great impact on increased prices. Higher numbers of reviews per month were found to be associated with a lower rental price.

Property type was found to be an important factor in predicting the availability of a property over the following 30 days, with houses being three times and serviced apartments being twice as likely to have some availability than apartments. As well as this, properties accommodating large groups of nine or more were twice as likely to have availability than properties accommodating fewer than nine. Further, properties owned by hosts with more than five listings were almost twice as likely to have availability.

Two main factors have been identified as leading to the property's very high review rating; the neighbourhood where the Airbnb is located, and the price per guest. [6] Barra da Tijuca is an upper-class neighbourhood, and it is one of the most developed places in Brazil. Therefore, the proportion of Airbnb in Barra da Tijuca with a very high score is substantially greater than in the other two districts, especially Copacabana. Airbnb with a higher price per guest is more likely to have an extremely high review rating.

Introduction

Airbnb describes the key factors to promoting your property as price, availability and how attractive your listing is [7]. They also name factors such as how many booking requests you accept, host response time and the option for *Instant Book.* These factors drive the listing rank algorithm on the Airbnb website, so it is expected that they would be correlated with availability.

The objective of this report is to identify the determinants of Airbnb accommodation prices and high review scores. We are also interested in predicting whether a listed property will be available for at least one night within the following 30 days. Our data contains information on the properties listed on the site in April 2018 for Rio de Janeiro, Brazil. The analysis will concentrate on the three districts with the most listed properties and the prevalent types of them.

Preliminary Analysis

There were 4583 observations in the dataset but only very few missing data points in general, with the exception to this being the host response rate (538 NAs / ~12% missing). This was found to be MAR, conditional on reviews per month, meaning inclusion of host response rate in the models suggests inclusion of reviews per month too to avoid bias. Some data required reformatting to carry out the analysis, for instance conversion to factor or numerical variables.

Extreme values were found for minimum nights observations. Observations over 30 days were marked as NA as we found no outside evidence that values over this were used in Airbnb properties. Similarly, bathroom observations of 0 were marked as NA, due to legal constraints of properties. Many of the continuous variables were highly skewed (Figure 1), so we had to use appropriate transformations or categorisations where possible.

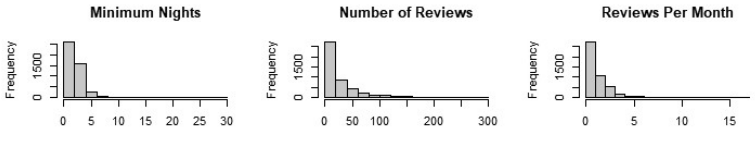


Figure 1: Continous variable histograms

Significant Determinants for Rental Price

The response variable price has a heavy upper tail which motivates us to consider the price on a log scale. The histogram is approximately symmetric after the log transformation. In addition, the choice of using the log transformation is very close to the optimum value (-0.10 vs 0) suggested by the Box-Cox transformation.

Price was plotted by the explanatory variables to evaluate if there was a transformation needed to achieve normality. Consequently, we use the log transformation on minimum nights, reviews per month and number of reviews. We also apply the square root transformation on guests included and square transformation on host response rate.

The inclusion of highly correlated variables in the model may result in collinearity. The number of guests included in the price and the number of accommodates are found to be highly correlated (0.49). The correlation between the number of reviews and the number of reviews per month is likewise strong (0.57). Adding each of them to a linear model at a time, then comparing the AIC, the AIC is lowest in the linear model with the number of guests included in the price and reviews. Neither adding an offset nor attempting to use a Gamma generalised linear model could improve this model any further.

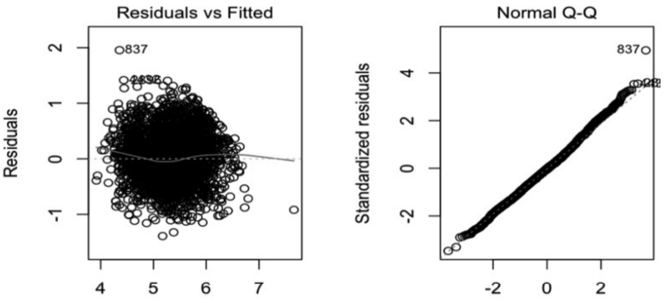
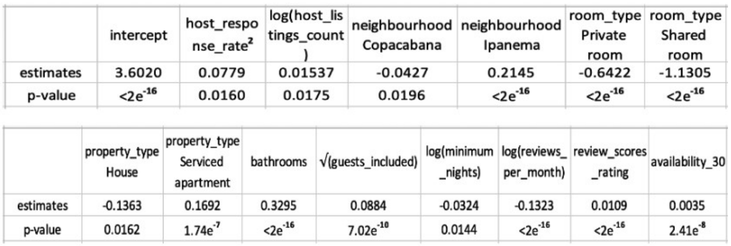


Figure 2 - Diagnostic Plots

Plotting a Q-Q plot of the model residuals against the normal distribution (figure 2), we see that they strongly follow the normal distribution and plotting the residuals against the fitted values there is no apparent relationship, suggesting good fit of the model. There was however an outlier (837) in the model, but after performing sensitivity analysis excluding this observation, it was deemed not to be an influential point. The proposed linear model was the following:

$$\begin{aligned} log(price_i) = \beta_0 + \beta_1 property\_type_i + \beta_2 room\_type_i + \beta_3 \sqrt{guests_i} + \beta_4 accommodates_i + \beta_5 log(listings_i) + \beta_6 log(reviews\_pm)_i \\ + \beta_7 host\_response_i^2 + \beta_8 bathrooms_i + \beta_9 neighbourhood_i + \beta_{10} availability_i + \beta_{11} rating_i + \epsilon_i \end{aligned} \quad (1)$$



	intercept	host_response_rate^2	log(host_listings_count)	neighbourhood Copacabana	neighbourhood Ipanema	room_type Private room	room_type Shared room
estimates	3.6020	0.0779	0.01537	-0.0427	0.2145	-0.6422	-1.1305
p-value	<2e^{-16}	0.0160	0.0175	0.0196	<2e^{-16}	<2e^{-16}	<2e^{-16}

	property_type House	property_type Serviced apartment	bathrooms	√(guests_included)	log(minimum_nights)	log(reviews_per_month)	review_scores_rating	availability_30
estimates	-0.1363	0.1692	0.3295	0.0884	-0.0324	-0.1323	0.0109	0.0035
p-value	0.0162	1.74e^{-7}	<2e^{-16}	7.02e^{-10}	0.0144	<2e^{-16}	<2e^{-16}	2.41e^{-8}

Table 1 - Coefficients and P-values of the Model

Exploring the relationship between the factors and the rental price (table 1), the variables with the largest effect were determined by fitting a scaled model, with the three most influential variables being the following: The price for properties in Ipanema were associated with a 24% [19%,29%] increase than those in Barra da Tijuca. The private room is associated with a 47% [45%,49%] decrease and shared room 67% [63%,72%] decrease in the price compared to the apartment. An additional bathroom was associated with a price increase of 39% [36%,42%]. Properties accommodating more guests were found to be related to higher prices. We estimated that a percentage point increase in the square root of number of people included in the price corresponds to an increase of 9% [6%,12%]. An additional review per month was associated with a price decrease of 13% [12%,15%]. An increase in review score rating of 1 lead to a 1.2% [0.8%,1.4%] increase in price.

Predicting Whether a Property Will Be Available for At Least 1 Night Over Next 30 Days

First, a binary variable was created describing whether or not there was at least 1 night available at the property over the next 30 days. Exploratory plots were produced to investigate any association between this availability variable and the other variables in our dataset. Property type and room type appeared to have an effect on probability of availability, with non-serviced apartments having lower availability than houses or serviced apartments, and shared rooms having higher availability than non-shared rooms.

Some continuous variables were categorised. This is when a linear/monotonic relationship was not seen, but there was a meaningful cut off point (or points) where an effect could be captured. Properties that accommodate a high number of people (>= 9) seemed to have a different probability of availability than those that accommodate less (figure 3). The minimum number of nights you could book the Airbnb for had a similar effect, with a cut-off of <=3 / >3, as well as properties where hosts had a lot of listings with a cut-off of <5 / >=5.

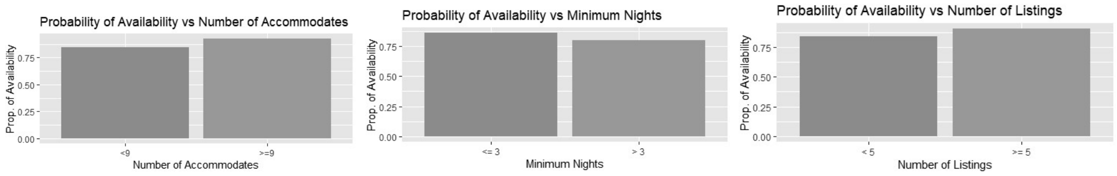


Figure 3: Probability of availability

The number of reviews as well as the reviews per month appeared to have an effect on the probability of availability, where in both cases an increase lead to a higher chance of availability of the property.

As the number of reviews and the number of reviews are highly correlated (0.4), by the fact that a higher number of reviews per month directly leads to a higher number of reviews, two binomial models were produced for 1 night availability (*Av*) of the following form:

$$Av_i \sim Bin(1, \mu_i)$$

$$log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 property\_type_i + \beta_2 room\_type_i + \beta_3 min\_nights_i + \beta_4 accommodates_i + \beta_5 listings_i + \beta_6 n\_reviews_i + \epsilon_i \quad (2)$$

$$log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 property\_type_i + \beta_2 room\_type_i + \beta_3 min\_nights_i + \beta_4 accommodates_i + \beta_5 listings_i + \beta_6 reviews\_pm_i + \epsilon_i \quad (3)$$

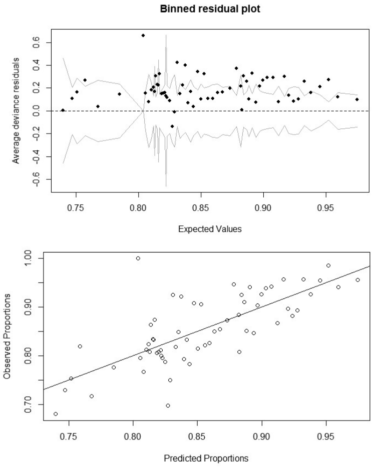


Figure 4 – Diagnostic Plots

Both models had similar AIC, and no relationship was seen in the deviance residuals, other than the models consistently over-predicting (possibly due to low FALSE responses in the data), suggesting the linearity assumption holds (figure 4). Further, the plots of observed vs predicted values had a linear and 1-to-1 relationship, with the only obvious signs of poorer fit being at particularly low proportions of availability.

In terms of predictive power, model 2 and 3 had a similar prediction error, using both $(Observed\ Response - Predicted\ Probability)^2$ and leave-one-out cross validation. The error was 563.7, vs an error of 683 if we were to simply predict TRUE every time, suggesting we have some predictive power but extra variables may improve this further.

While the model diagnostics and predictive power suggest good fit for both models, we do have a contextual reason to believe that in model (3) multicollinearity may exist. It is likely that a relationship exists between the minimum nights and the reviews per month, with higher minimum nights leading to less visits per month and therefore less reviews. For this reason, we instead favour model (2), with the diagnostic plots shown in Figure 4.

Interpreting the model, we find that houses are associated with being 3.37 [1.04,10.9] times as likely to have at least 1 night's availability, and serviced apartments are 2.64 [1.42,4.91] times as likely, when compared to non-serviced apartments. Further, properties with a minimum night's requirement of more than 3 nights are 0.680 [0.549,0.841] times as likely to have availability than those with less than this. Properties that accommodate a total of 9 or more people are 2.33 [1.21,4.47] times as likely than properties that accommodate less. Properties owned by hosts that have at least 5 listings are 1.74 [1.38,2.19] times as likely than those owned by hosts with less. An increase in number of reviews of a property of 1 lead to a 1.0% [0.6%,1.4%] increase in the probability of availability. Room type was found not to be a significant variable in the model, suggesting it does not affect the probability of availability.

The effect that number of reviews has on availability may be explained by average stay length. Airbnb's with more reviews likely have a shorter average stay length, meaning more gaps between guests, possibly increasing availability likelihood. "Superhost" status pushes properties to the top of the search results. This likely affects availability, as the listing will be viewed more, so bookings are more likely to be filled. Variables that affect superhost status include response rate and average review rating, which we have. It also includes, number/length of stay, and host cancellation rate. Host cancellation rate could be an instrumental variable on number of reviews, with higher cancellations leading to a lower number of reviews, however being independent on availability. These variables would be useful to have for our analysis.

Factors that Lead to the Property Receiving Very High Review Scores

We look to analyse the variables associated with extremely high review scores in Brazil's Airbnb data. Firstly, 4.8 out of 5 is considered a high rating [3] or equivalently 96 out of 100. A binary review score variable was established to describe whether the review score rating is greater than 96.

Explanatory plots were produced to find the relationship between very high review scores and other variables. Probability of very high review scores was found to be influenced by the neighbourhood and room types. Bathrooms, bathrooms per guests, price, price per guests, reviews per month and host listings count were found to have a relationship with very high review scores. There is a 0.56 correlation between bathrooms and price, suggesting the variables are collinear. Therefore, further analysis will focus on the number of bathrooms per guest and the price per guest instead. The effect of availability appears to be quadratic; a square term is included in the model. The final model we propose is a binomial model to model the binary review scores rating as the response variable.

By using a 2-way ANOVA test, we discovered two interaction terms: one between reviews per month and host listings count, and the other between reviews per month and availability in 30 days.

We determined our best model to be:

$$Review\_Score_i \sim Bin(1, \nu_i)$$

$$\begin{aligned} log(\frac{\nu_i}{1-\nu_i}) = \beta_0 + \beta_1 \frac{price\_per\_guest_i}{10} + \beta_2 bathrooms\_per\_guest_i + \beta_3 reviews\_pm_i + \beta_4 neighbourhood_i + \beta_5 room\_type_i \\ + \beta_6 listings_i + \beta_7 availability_i + \beta_8 availability_i^2 + \beta_9 reviews\_pm_i\!: listings_i + \beta_{10} reviews\_pm_i\!: availability_i + \epsilon_i \end{aligned} \tag{4}$$

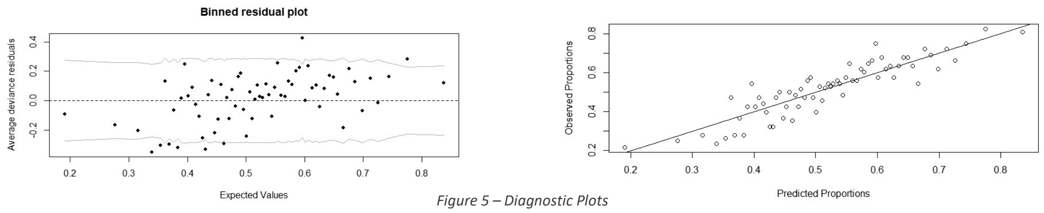


Figure 5 – Diagnostic Plots

The diagnostic plots (figure 5) look reasonable, most of the dots are in the confidence bands, with some under-prediction for low observed values, suggesting slightly poorer fit at this tail. Coefficients for the model are given below. (table 2)

The top three most significant variables are neighbourhood Copacabana, reviews per month and price per guest. However, reviews per month and high review scores are likely to instigate each other.[4] If the review score is high, it will attract more guests by the "Superhost" status, resulting in more guests and so more reviews per month. More reviews per month could suggest that the Airbnb is popular, and the rating will be relatively high if the host is consistent. Then only 2 most significant variables were left, neighbourhood and price per guest. The proportion of very high review scores was found to be 0.572 [0.478,0.685] times lower for Airbnb in Copacabana compared to Barra da Tijuca. A 10 dollar increase in the price per guest was found to be associated with a 1.376% [0.875%,1.881%] increase in the probability of very high review scores rating when every other variable is fixed in the model.

	intercept	Neighbourhood Copacabana	Neighbourhood Ipanema	Availability 30	Price per guest/10	Room type Private room
estimate	-0.138	-0.529	-0.356	0.0488	0.0119	0.0386
P-value	0.312	$2.78e^{-10}$	0.000313	0.000108	$1.47e^{-7}$	$7.88e^{-5}$

	Room type Shared room	Reviews per month	Listings count	Bathrooms per guest	availability^2	Reviews per month: listings count	Reviews per month: availability30
estimate	-0.288	0.438	-0.0168	0.348	-0.00178	-0.0135	-0.00951
P-value	0.402	$8.40e^{-10}$	0.0259	0.0249	$4.93e^{-6}$	0.00974	0.00484

Table 2: Estimated coefficients and p-values for model (4)

Availability and host listing count are more likely to be confounders because they have association with neighbourhood. Availability has a positive effect on the probability of extremely high review score. Airbnb in Barra da Tijuca has the highest proportion of very high review scores and the greatest number of available nights compared to Copacabana and Ipanema as shown by the box plots (figure 6). Equivalently, listing count has a negative coefficient in the model and hosts who have few properties are more likely to be in Barra da Tijuca (figure 6). Therefore, availability and host listings count appear to be confounders. As we found in Q1, room type and bathrooms are strongly associated with price, which suggested room type and bathrooms are confounding factors too.

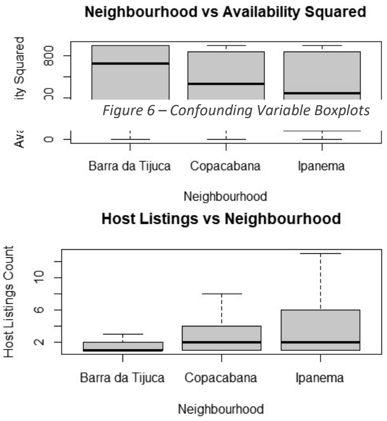


Figure 6 – Confounding Variable Boxplots

Data Limitations & Conclusion

The data provided was from three districts of Rio and perhaps are not representative of all properties on Airbnb. In terms of generalisability, this means that our results on price predictions may not be able to be extended to other regions, as we found neighbourhood was a significant predictor. In other regions there may be a larger disparity in the neighbourhoods leading to greater price differences. On the other hand, the variables used to predict availability were more consistent with the global averages. For instance, the majority of review scores were high and there were few properties (1.44%) with an average review score of less than 80. However, this is common for Airbnb listings. A study on 1.3 million properties spread across the globe found that 98% of properties scored at least 4 out of 5 stars [3].. This means that these results would be generalisable to other properties around the world.

The data is temporally limited as only one time period is considered; therefore, seasonal differences are missing. It would be expected that more popular tourist areas would observe a sharp increase in price during the busiest periods due to high demand – for instance, the Easter holiday. Further, it is not clear whether all observations are independent, as it's possible some come from the same hosts. This is suggested in the data by the presence of duplicate host listing counts and host response rates. Using these two variables in our analysis may therefore introduce bias and so it would have been useful to have a "Host ID" variable to control this.

The likelihood of very high review scores is directly related to the price per guest. However, excellent service is more likely to be the cause of the high review scores rating [3], and it is related to the pricing. It's also possible that the price is a confounding factor. However, there are no variables in this data set that are relevant to the service.

We found that the neighbourhood, the type of room, the number of bathrooms and the number of guests included significantly influence the property price. Greater accommodation capacity is associated with higher prices, as is the provision of more bathrooms. The number of reviews per month negatively affects the Airbnb price. A possible explanation is that the quantity of reviews is explained by demand and there is more demand for less expensive listings[2]. In terms of lower price correlating with lower ratings, we found a study that the overall distribution of the data shows that there are an overwhelming number of higher scores. This means that reviews for properties are likely to be quite high[5].

We found that property type, number of host listings, minimum night requirements, number of reviews, and the number of guests that an Airbnb property accommodates affects the probability of the property having at least 1 night available over the following 30 days. The property type had the greatest relative effect on the availability, followed by the number of guests, host listings, number of reviews and finally the minimum nights (which had the smallest relative effect).

Neighbourhood and price per guest lead to properties receiving a very high review score. Neighbourhood is the most significant reason. The probability of very high review scores in Barra da Tijuca is much higher than Copacabana and Ipanema, especially Copacabana. Price per guest and reviews per month had a positive effect on extremely high review scores. All other variables defined before either have no significant effect or they confound or instigate others.

Bibliography

[1] THE zebra. 2022. Airbnb statistics and host insights [2022]. [online] Available at: <https://www.thezebra.com/resources/home/airbnb-statistics/> [Accessed 28 April 2022].

[2] Perez-Sanchez, V., Serrano-Estrada, Marti and Mora-Garcia, 2022. The What, Where, and Why of Airbnb Price Determinants. [online] Ideas.repec.org. Available at: <https://ideas.repec.org/a/gam/jsusta/v10y2018i12p4596-d188018.html> [Accessed 30 April 2022].

[3] BnB Facts [online]. Available at https://bnbfacts.com/airbnb-ratings-overview/ [Accessed 1 May 2022].

[4] Medium [online] Available at https://medium.com/@agustinus.thehub/this-is-how-you-get-more-bookings-and-higher-ratings-as-airbnb-host-ff2846834b3e [Accessed 11 Sep 2018].

[5] Medium. 2022. Factors to increase rental prices for AirBnB. [online] Available at: <https://elibrunette.medium.com/factors-to-increase-rental-prices-for-airbnb-6a4cbb928e0d> [Accessed 2 May 2022].

[6] Barra da Tijuca-Wikipedia [online] Available at: https://en.wikipedia.org/wiki/Barra_da_Tijuca [Accessed 3 May 2022].

[7] Airbnb [online]. Available at https://www.airbnb.co.uk/resources/hosting-homes/a/how-airbnb-search-works-44 [Accessed 3 May 2022].