Analysis of AirBnb Listings and Predictors of Price

Anuraag Kumar

***Abstract*— The paper uses data on AirBnb listing and reviews with Boston from the years 2016-2017 to attempt to create an explanatory model for what causes the wide variation in listing price. We used apriori algorithml, PCA dimension reduction, and sensitivity analysis to conclude that neither the listing's room specifics nor the content of the reviews of each listing could explain the large variation in prices.**

I. INTRODUCTION

Vacations are the things that fill the dreams of bored office workers and fill the Instagram profiles of the people I'm jealous of. Travel is a multi-billion dollar industry, and hotels are sometimes too expensive, too out of the way, and too out of touch. This is where Air Bread and Breakfast comes in. AirBnB is a website and an app that controls booking for a self-run bread and breakfasts across the nation. AirBNB has been in operation since 2008 and was founded by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia after they started a Bread and Breakfast in their own living room to help pay for rent.

One key feature of using AirBnb is the review system. I predicted that there would be a equal amount of listings with negative and positive reviews. This hypothesis was very wrong as the vast majority of reviews were positive.

I also expected most AirBnB reviews scores to be correlated with the listed price. This was very rarely the case. There was a very low correlation between how well something was reviewed and how expensive the listing was.

II. DATA

We used data scraped directly from AirBnB itself from the September 2016 to September 2017 specifically in the city of Boston. This resulted in three datasets. We first has listings which had a listing ID and many variable associated with the listing as a whole like host response rate, description, number of beds, number of baths, full capacity, location and much more. The listing dataset was by far the largest dataset by far with the vast number of features. The dataset contained 3585 listings across the city of Boston. That number is surely much larger now. Also, the mean listing price was 173 dollars with a median of 150 dollars and a standard deviation of 140 dollars. This is an incredibly large level of deviation amongst the price among a specific location. This implies a lack of control by AirBnB in controlling the price during this period. It might also be due to a wide variety of different listings in Boston. Speaking of wide variety, i t might interest to you know that there were listings for as low as 10 dollars a night during this time. Also, many listings had the default setting for maximum nights. This meant that somebody could live in a specific AirBnB place for 1125 nights. This dataset also had a lot of gaps in the reviews per month column. This implied bad data collection methods, as this feature isn't hard to track per listing. In general, this problem extended to the rest of the dataset as there are many many missing values. I didn't analyze those columns in general. One would be hard pressed to find a row that didn't contain at least one missing value. However, many of these columns were not important in the analysis so they were not important in our analysis.

There was another dataset called reviews which contained reviews for each listing at different points in time. There are a total of 68,785 reviews which implies that there is an average of 19 reviews per listing over the course of the year. This means that there was an average of just over 1 reviewing person staying in each listing every 3 weeks. The reviews data sheet contains the date the review was made and the specific comments the reviewer made.

The last dataset was calendar which marked down every day whether or not a listing was booked. This dataset had the least utility in our project. It did have the most rows though at a whopping 1,308,890 different data points.

III. RESULTS

We start our analysis by first considering what kinds of listings are the most frequent kind of listings. We used the apriori algorithm here, and I used the following columns as differentiating factors between different listings.

- Bedrooms
- Bathrooms
- Maximum Capacity
- Room Type
- Property Type

The apriori algorithm sorts different itemsets by classifying them as frequent or infrequent. I ran the algorithm using a hand coded method and also an imported method and I got the same answer in both cases. Nevertheless, here are the 5 most frequent itemsets.

- (1 bed), Support = 0.77
- (Apartment), Support = 0.73
- (1 bed), Support = 0.66
- (1 bed, Apartment), Support = 0.60
- (Entire home/Apartment), Support = 0.59

The different between Apartment and Entire Home/Apartment is that the first one was classified as property type and the second one was a room type. The second one offers the entire property as a listing instead of just a singular room or a couple of rooms.

We also have a listing of less frequent itemsets that were contained to minimum support values of 0.1 and 0.2. For 0.1,

- (2 bed, 1 bath, Entire Home/Apartment), Support = 0.100
- (1 bath, 2 bed), Support = 0.100
- (2 baths, Entire Home/Apartment), Support = 0.101
- (1 bed, Total Capacity of 1, Private Room), Support = 0.101
- (Total Capacity of 1, Private Room), Support = 0.101

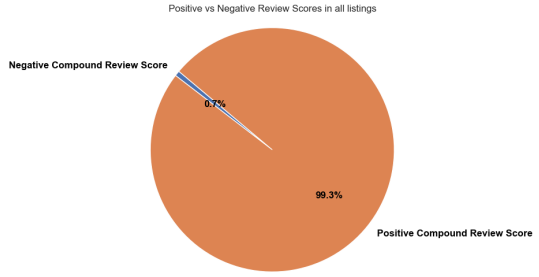
and for 0.2

- (1 bed, 1 bath, Total Capacity of 2, Apartment), Support = 0.216
- (1 bed, 1 bath, Apartment, Entire Apartment/House), Support = 0.217
- (1 bed, Private Room, Apartment), Support = 0.219
- (Private Room, Apartment), Support = 0.219
- (1 bed, Apartment, Entire Apartment/House), Support = 0.22515

We find a couple of very clear association rules from this data. The first is that Private Rooms are very likely to have only 1 bed. We can see this association rule in both lists. The second important finding is that 1 bed is very common. Since 1 bed is so common, it doesn't really account for the wide variety of prices. 1 bed and apartment accounts for over 50% of all listings. How is it then that the prices vary so wildly? It is clear that looking at itemsets won't give us the answer.

I began looking at multilinear regression as a possible method. I hypothesized earlier that quality of reviews could be associated with price. To rate the reviews that came in, I started doing sentiment analysis on the reviews. To start with a simple litmus check, I used two lists of positive and negative words to see what proportion of each review's words were positive and how many were negative. I found that most reviews skewed positive. Amongst the reviews, many didn't have a single negative word in the review.

To make sure I was getting accurate results, I then used a more sophisticated method by using nltk package on python to help me. I found that there was something very similar going here as well. The vast majority of reviews were positive. In the sentiment analysis, I calculated a composite score which just subtracted the negative score from the positive score. Here are the results visualized.



We can see how positive the score are here. We can conclude that most people are happy with the services. It seems that we don't have a way of concluding price from here. Therefore, we must include more variables in our explanatory model of price. For this paper, I went with the following

- host response rate
- review scores rating
- review scores accuracy
- review scores cleanliness
- review scores check-in
- review scores communication
- positivity mean
- negativity mean
- positivity simple mean
- negativity simple mean

Where the positivity simple mean and negativity simple mean were the results of averaging the review scores I got from my simple regression. After running the regression, I got the following results.

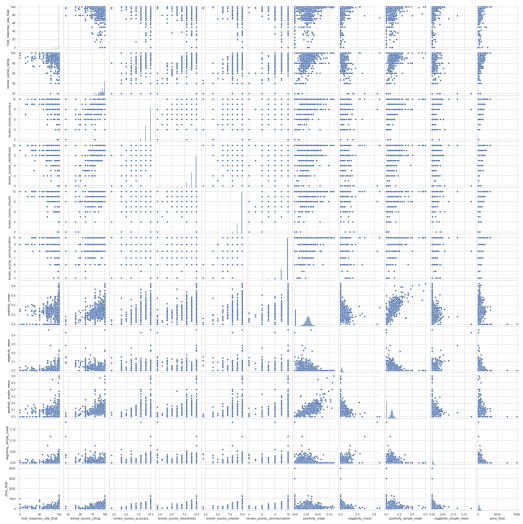
OLS Regression Results

Dep. Variable:	price_float	R-squared:	0.027
Model:	OLS	Adj. R-squared:	0.023
Method:	Least Squares	F-statistic:	7.808
Date:	Tue, 17 Oct 2023	Prob (F-statistic):	1.82e-12
Time:	01:37:20	Log-Likelihood:	-18143.
No. Observations:	2868	AIC:	3.631e+04
Df Residuals:	2857	BIC:	3.637e+04
Df Model:	10		
Covariance Type:	nonrobust		

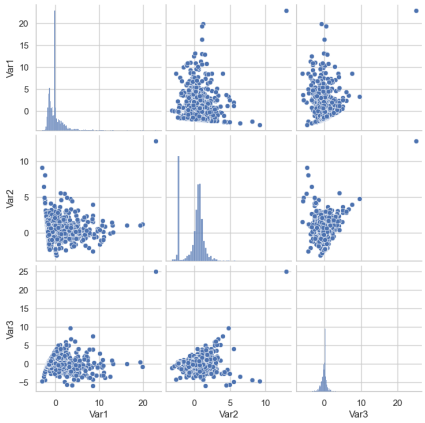
	coef	std err	t	P>|t|	[0.025	0.975]
const	193.1397	48.015	4.022	0.000	98.992	287.288
x1	-0.0257	0.215	-0.120	0.905	-0.446	0.395
x2	1.6380	0.600	2.730	0.006	0.461	2.814
x3	-11.9449	4.516	-2.645	0.008	-20.799	-3.090
x4	16.8282	3.915	4.299	0.000	9.152	24.504
x5	-13.2424	5.294	-2.502	0.012	-23.622	-2.863
x6	-6.7480	5.677	-1.189	0.235	-17.880	4.384
x7	-159.0000	41.713	-3.812	0.000	-240.790	-77.209
x8	-209.3744	182.760	-1.146	0.252	-567.730	148.981
x9	253.6703	113.428	2.236	0.025	31.261	476.080
x10	528.6053	559.771	0.944	0.345	-568.991	1626.201

Omnibus:	3279.420	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	736005.532
Skew:	5.526	Prob(JB):	0.00
Kurtosis:	80.698	Cond. No.	3.02e+04

We find that this model is a bad predictor of price changes as $R^2 = 0.027$ which implies a very weak correlation. The standard error on the coefficients is also very high which implies a weak and statistically insignificant correlation. It seems that our effect size was small. To further analyze, the coefficients are listed in the same order I listed them above. Therefore, we have that our sensitivity analysis had nothing to do with the price at all as the 95% confidence interval contained 0 for all of those values. It seems that the review scores for rating, accuracy, cleanliness, and check-in were all significant as well as host-response rate. This makes sense as the more one pays, the more likely the owner of the property will have a greater stake in the building. There was also a small amount of colineraity between variables as shown by the following pairplot.



One can see that a couple variables are related to each linearly. I tried then to reduce the amount of dimensions. It was clear that the review score variables and the sensitivity analysis variables were related to one another. For that reason, I tried to reduce the dimension of our explanatory set to 3 (there are 3 distinct category of variables), so perhaps that would result in a better regression. After trying that I got the following regression.



We have low correlation between the variables, which is good. Reducing dimension on poorly chosen explanatory variables doesn't make problem better. It seems that we forgot to include where in Boston our listing was located. It also seems we forgot to include cleaning and extra costs into our analysis as that would move things around. Nevertheless, it seem we have a problematic situation with AirBnB in Boston. Reviews cannot explain the large variation in price and neither could type of listings. It makes one wonder what was the reason for this large variation. This could be explored later in future studies.

OLS Regression Results

Dep. Variable:	price_float	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	6.419
Date:	Tue, 17 Oct 2023	Prob (F-statistic):	0.000249
Time:	01:34:01	Log-Likelihood:	-18172.
No. Observations:	2868	AIC:	3.635e+04
Df Residuals:	2864	BIC:	3.638e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	173.7385	2.552	68.067	0.000	168.734	178.743
x1	-3.4263	1.300	-2.636	0.008	-5.975	-0.878
x2	-6.1830	1.781	-3.471	0.001	-9.676	-2.691
x3	-1.2264	2.432	-0.504	0.614	-5.995	3.542

Omnibus:	3232.817	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	686970.349
Skew:	5.400	Prob(JB):	0.00
Kurtosis:	78.047	Cond. No.	1.96

Here we get an even lower R^2 value. Also, our third variable is not significant. This is an indication that lowering our dimension using PCA was a bad idea. We can show our pairplot for the variables in this regression to demonstrate the problem with this approach.