Full-Time Airbnb:
What Can You Make With Your Home?

Jon Whelpley

December 2020

Contents

1 Introduction

Airbnb has been a popular alternative to hotels for any person interested in travelling. Moreover, it has given homeowners an opportunity to supplement their income by leasing out a spare bedroom or spending a few nights away from the home altogether. Naturally this lends itself to the question: What revenue can you realistically make if you were to lease a home as a full time airbnb?

2 Data

2.1 The Main Data Set

The data comes in the form of a table taken from Kaggle. Here is the Link
The table consists of 226030 records with 17 features which are as follows:

Column Name	Description
'id'	unique listing id
'name'	name/description of listing
'host_id'	unique host Id
'host_name'	name of host
'neighbourhood_group'	group in which the neighbourhood lies
'neighbourhood'	name of the neighbourhood
'latitude'	latitude of listing
'longitude'	longitude of listing
'room_type'	room type
'price'	price of listing per night
'minimum_nights'	min no. of nights required to book.
'number_of_reviews'	total number of reviews on listing
'last_review'	date of last review
'reviews_per_month'	average reviews per month on listing
'calculated_host_listings_count'	total number of listings by host
'availability_365'	num of days in year listing is available
'city'	region of the listing

The goal here is to predict the reviews per month. According to airbnb, users leave a review 70 percent of the time. This will give us an estimate for the number of consumers that actually rent the bnb. For example if we predict a home will have 10 reviews per month at 80 dollars per night, and a minimum of 2 nights per stay we will estimate the monthly revenue to be:

$$\frac{10 \text{ reviews per month}}{0.7 \text{ reviews per stay}} * \text{ 2 nights per stay} * 80 \text{ dollars per night}$$

$$= 2285.7 \text{ dollars per month}$$

Essentially, if we're able to predict reviews per month based on these features we can give a projection of revenue and even compare that to mortgage and maintenance costs.

2.2 Additional Location Data

While Location data for each airbnb is already given. Travellers don't just want to get an airbnb to have a place to stay. They travel to find things to do. So it will be very important to lookup places where there's interesting things to do. Due to the limitations of four-square's API it would be disadvantageous to check each airbnb and find what's nearby; however, if we cluster the neighborhoods we may find that searching the cluster centers may give enough to information to find what's close by. Minimally this will be used for the visualizations but it might be possible to get features that are just as valuable by getting locations that are nearby the centers of neighborhood cluster in a certain area.

2.3 Descriptive Data

Not only will extra location information be important but I suspect the text in the 'name' column will be important as it describes and 'sells' the airbnb. I will be looking for the most common words through the listings and see if those influence the number of reviews in any way. Alternatively I may even try to find the most common words to the top 10 percent of the most reviewed so showcase which descriptions are most successful. Each word will then become its own feature column that is '1' when the name column has the word and '0' when it doesn't.

3 Process and Methodology

3.1 Cleaning

The processes of cleaning the data can be sorted into the following steps:

1. Drop any unnecessary column information
2. Create dummy variables for categorical data
3. Find a method for numerically describing the name column
4. Focus in on a single city for preliminary modelling

The features I chose to ignore from the original data frame were the 'id', 'host_id', 'host_name', 'neighborhood_group', 'neighborhood', and 'last_review'. the listing id, host, id, and host name columns are just nomenclature and it seems irrelevant to actually predicting the success of an airbnb. I ignored the neighborhood group and neighborhood columns because there already exists exact locations and the cities with those locations. The last review column was dropped because date-time data would've added some clutter to the data more than it would've been relevant.

'room_type' is the only categorical variable besides 'name' and 'city' that seemed significant. There were also only 4 categories within the column so making it a dummy variable and adding 4 columns wouldn't make things too computationally expensive later.

For the 'name' column the words in each row were tallied and totaled into a python dictionary. Basically, the more often the word appeared in the name column, the higher the tally. This gave a sense of what keywords were most popular in the listings. So about 100 columns were added with keywords. In hindsight, it may even be beneficial to find the most common words for the listings with the highest number of reviews but for now we'll use the popular words as a benchmark.

After all of this, the data was parsed into several tables by city. This may it would be much easier to focus in on smaller data as opposed to a 200K+ record data table. Not to mention if there was any further data manipulation necessary, it'd be more efficient and it could be tailored to each city specifically because trying to generalize each airbnb by strictly latitude and longitude is prone to significant under fitting.

3.2 Collecting Location Data

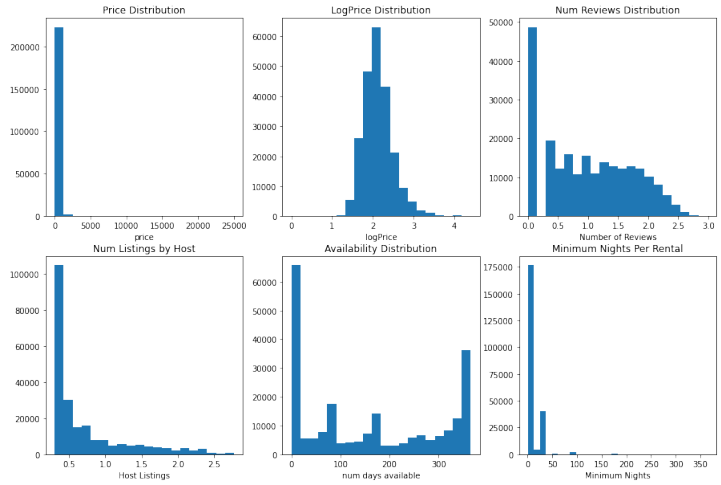
Before using four square's API, I had to be conservative about how I collected location data of nearby social spots. The reason being that I had a lot of data and a limited number of calls. So instead, I decided to divide the city into location clusters and use the cluster centers for the search. Then I joined the search results to give a list of locations for cafes and restaurants. Then I counted the number of locations within a "square mile" (this was approximated with the

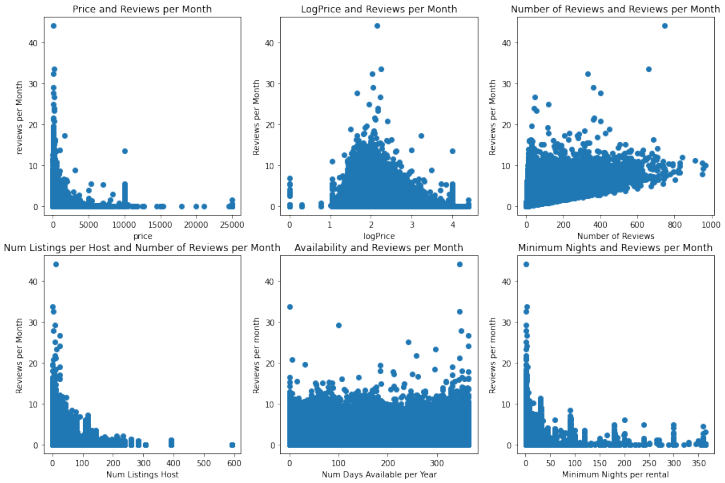
latitude and longitude) of each airbnb within the city to give a new feature for each kind of search. These were 'nearby_cafe', or 'nearby_restaurant'. It could really be anything worth searching for.

Given enough real estate savvy and a little keyword optimization, one may actually be able to find the most relevant nearby city attractions and activities that could being in more people. But for a preliminary model cafes, bars, and restaurants will be a good start.

3.3 Visualizing

To visualize the data and see how each data point influences the target variable we want to predict an array of histograms and scatter plots were made. Notice that because of some outliers there's a lot of heavily skewed data. To combat this, 'logPrice' was included to 'normalize' the price which is measured on a logarithmic scale (base 10). The data from the scatter plots are messy and it seems that any single variable is not sufficient to predict the success of an airbnb just from an initial look. That being said. We can see inverse relationships between price and reviews; however, prices can be too low to where it is detrimental to the owner to make renting their bnb as cheap as possible. And naturally, if you require longer stays, you'll have less reviews per month, if you have one person stay for 1 month that only 1 review but you also maximize your revenue for that month.

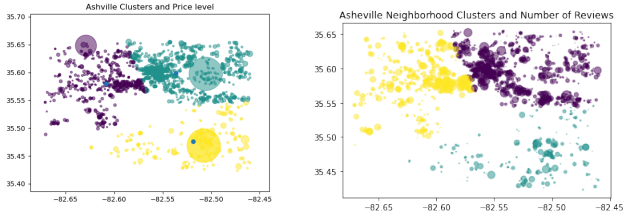




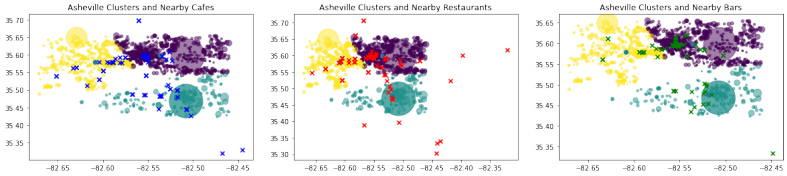
One important note to make is that the words in the listing might be able to give some detailed insight into the qualitative aspect of an airbnb. Here's a word cloud to illustrate which words are most popular for each listing.

3.4 Focusing on a Single City

So in order to best visualize how location affects prices and reviews it'd be nice to be able to look at price and location. The data is centered around very specific cities and so it might be beneficial to make city specific models. It also helps in trying to find nearby cities and cafes for each location to see how much value is added by being near those spots. The following scatter plots show latitude and longitude with color coded neighborhood clusters and the size of each circle is proportional to price and number of reviews per month.



Additionally here are some maps that show locations of cafes, restaurants, and bars respectively:



3.5 Modeling

Of course since we're predicting a continuous variable, that means regression models are the best tools for the job. A variety of regression techniques starting with standard multiple regression and trying ridge, lasso, and polynomial (degree 2) regression with an array of hyper parameters to fine tune the model. All of them performed pretty similarly for the single city of Asheville aside from polynomial regression which instead over fit the model which is to be expected considering that with over 100 features and about 2000 records, the polynomial features would outnumber the sample size. All the models tested performed similarly in terms of their r2 scores regardless of hyper parameter tuning; however, the addition of word and location features did do something to improve the model (although location features did improve the model negligibly) when they were applied to a particular city. Given more time and a better pipeline it would have been beneficial to do a different regression model for each city as opposed to just applying one size fits all regression model.

4 Results

This is how the initial regression tests went in making the various predictive models.

4.1 The Data Overall vs. A Particular City

Data	Model	R2 Train	R2 Test
All Cities W/out Word Data	Linear Model	0.58431	0.59121
All Cities W/out Word Data	Polynomial Model	0.64422	0.63856
All Cities W/ Word Data	Linear Model	0.60066	0.60312
Asheville	Linear Model	0.48627	0.53488
Asheville W/ Word Data	Linear Model	0.69912	0.58116
Asheville W/ Word and Location Data	Linear Model	0.68902	0.62481
Asheville	Polynomial Model	0.55369	0.51873
Asheville W/ Word Data	Polynomial Model	0.99457	-5.46×10^{20}
Asheville W/ Word and Location Data	Polynomial Model	1.00000	-166711.89

4.2 Ridge and Lasso Regression for a Single City

Data	Model	alpha	R2 Train	R2 Test
Asheville W/ Word and Location Data	Ridge Model	0.001	0.67979	0.66250
Asheville W/ Word and Location Data	Ridge Model	0.003	0.67811	0.66660
Asheville W/ Word and Location Data	Ridge Model	0.01	0.70804	0.52569
Asheville W/ Word and Location Data	Ridge Model	0.03	0.68249	0.65147
Asheville W/ Word and Location Data	Ridge Model	0.1	0.68900	0.62419
Asheville W/ Word and Location Data	Ridge Model	0.3	0.68706	0.63987
Asheville W/ Word and Location Data	Ridge Model	1	0.69657	0.60516
Asheville W/ Word and Location Data	Ridge Model	3	0.69346	0.61418
Asheville W/ Word Data	Ridge Model	0.001	0.69436	0.41451
Asheville W/ Word Data	Ridge Model	0.003	0.68991	0.61929
Asheville W/ Word Data	Ridge Model	0.01	0.68737	0.62442
Asheville W/ Word Data	Ridge Model	0.03	0.68618	0.63737
Asheville W/ Word Data	Ridge Model	0.1	0.69309	0.60153
Asheville W/ Word Data	Ridge Model	0.3	0.69591	0.60583
Asheville W/ Word Data	Ridge Model	1	0.68610	0.63966
Asheville W/ Word Data	Ridge Model	3	0.68696	0.64244
Asheville W/ Word and Location Data	Lasso Model	0.001	0.67375	0.68329
Asheville W/ Word and Location Data	Lasso Model	0.003	0.67831	0.65359
Asheville W/ Word and Location Data	Lasso Model	0.01	0.66684	0.65547
Asheville W/ Word and Location Data	Lasso Model	0.03	0.66295	0.60428
Asheville W/ Word and Location Data	Lasso Model	0.1	0.64774	0.61475
Asheville W/ Word and Location Data	Lasso Model	0.3	0.63673	0.64630
Asheville W/ Word and Location Data	Lasso Model	1	0.64632	0.60976
Asheville W/ Word and Location Data	Lasso Model	3	0.63800	0.60862
Asheville W/ Word Data	Lasso Model	0.001	0.67579	0.67194
Asheville W/ Word Data	Lasso Model	0.003	0.67363	0.66844
Asheville W/ Word Data	Lasso Model	0.01	0.65099	0.69421
Asheville W/ Word Data	Lasso Model	0.03	0.65261	0.6421
Asheville W/ Word Data	Lasso Model	0.1	0.64426	0.61422
Asheville W/ Word Data	Lasso Model	0.3	0.64007	0.62722
Asheville W/ Word Data	Lasso Model	1	0.64304	-0.56528
Asheville W/ Word Data	Lasso Model	3	0.63863	0.61232

5 Other Considerations

This is to discuss any extra information that would be important for model refinement in case anyone wanted to try to replicate this project and refine it.

5.1 What the data doesn't cover

This is just a list of features that would probably be worth looking into:

1. Square footage
2. Style and Cleanliness of the space
3. Expenses to actually maintain a home
4. Ratings for each bnb listing
5. Adding customer keyword data i.e. "What are tenants looking for?"
6. Peacock factors like pools, outdoor space, private bowling alleys, or whatever make the home fun

5.2 What Could Be Technically Improved

The first thing that could definitely be improved is adding more descriptive data. The second is finding better location indicators that add to the value of an airbnb and using density based clustering on the locations of each airbnb to indicate a 'competition metric.' In other words, the 'denser' the cluster the more competitive the market space. It would have also been nice to determine which features are most predictive although the word vectors certainly did add to the accuracy of the model. Lastly, it would probably be ideal to get rid of any significant outliers because let's face it, if you're charging 25000 dollars per night to stay at your airbnb, you probably don't have time to look at some amateur data scientist's analysis of a very specific section of the "hospitality" market space.

If this were to be implemented into something more user friendly and less technically thrown together, there'd have to be some functions to streamline the data through an entire custom pipeline to 1. clean the data, 2. do the basic language processing, 3. Cluster the data to add appropriate location data and location dependent features, and 4. Shove everything into the regression pipelines and give a projection of monthly revenue.

6 Conclusion

6.1 For the Homeowner

This analysis shouldn't add to what you probably already know. If you're going to be out of town and you want to make a few extra bucks and don't mind cleaning up after some college students discovering who they are by travelling to Portland to try out every hipster coffee shop they can before moving back in with their parents, then yes, you should put your home on airbnb. If you're paying a mortgage anyways and you like to travel, this is the perfect side hustle.

6.2 For the Investor

I've got some good news and bad news. The bad news is that you are very unlikely to start seeing any profit from your airbnb business venture. If we go with strictly the kaggle data, you're only likely to see MAYBE a few customers. However, there are a few tips and takeaways that may work based on some toy samples made in the notebook:

1. Don't worry too much about price, but if you set the price too low it leads to fewer reviews
2. Longer stays mean more consistent cash flow.
3. Availability doesn't do much to influence the number of people that come through
4. It's about as profitable to rent out rooms individually as it is to rent out the entire home
5. Build as good a reputation as you can early on. The more reviews you have, the more consistent your revenue stream

At the end of the day this doesn't any more or less profitable that being a landlord and charging rent. Aside from the risk of being between tenants, it's still a more consistent and likely more secure investment if you have some real estate savvy.