Making Money with Airbnb

PREDICTING THE SUCCESS OF YOUR AIRBNB

“There are 3 kinds of lies:
lies, damned lies, and statistics.”

- LEONHARD EULER, 1706 (THIS IS ALSO A LIE)

The Question

Can we make money with Airbnb?

In other words:
Can we predict the number of customers an Airbnb will have each month and estimate the revenue per month?

What's the point of this?

1. To complete a Coursera capstone project

2. To make side hustle money while changing careers to data science through Airbnb

3. To give real estate investors some data driven insight into what they probably already know about housing and Airbnb

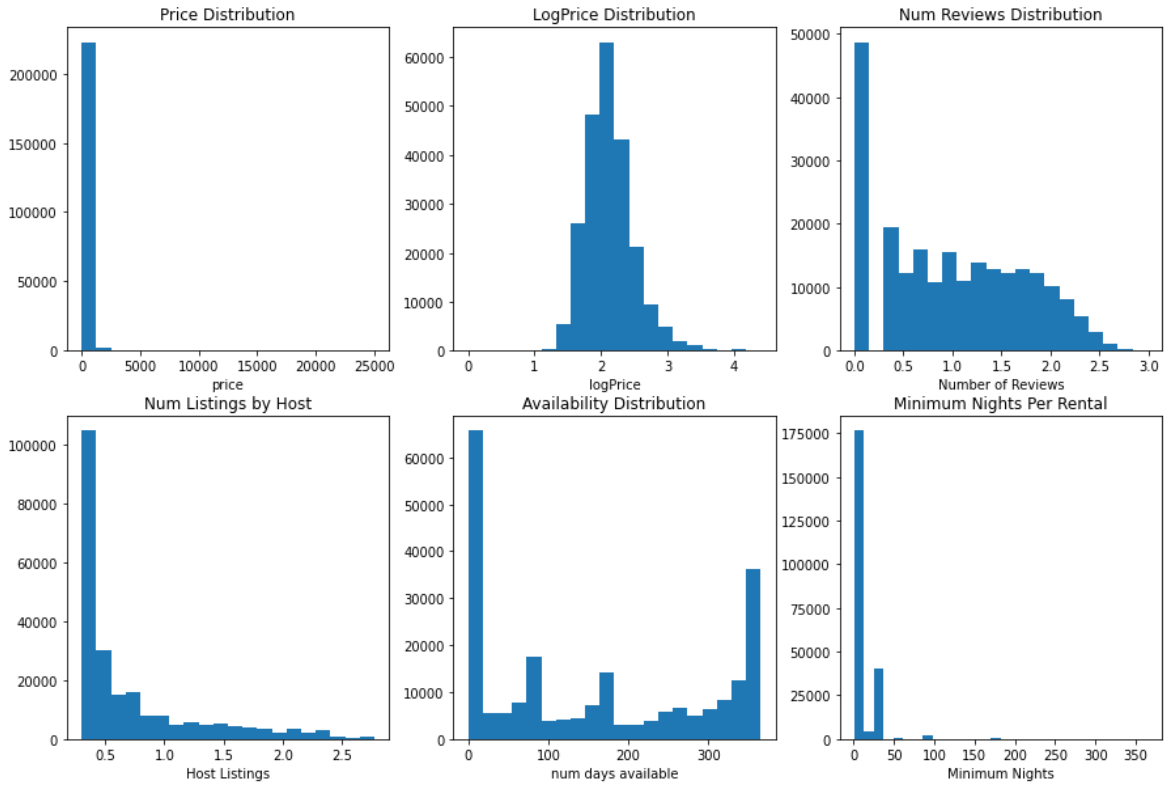
4. To have an excuse to play with computer toys

5. Because the data is freely available on Kaggle aside from a few datapoints we'll get to later

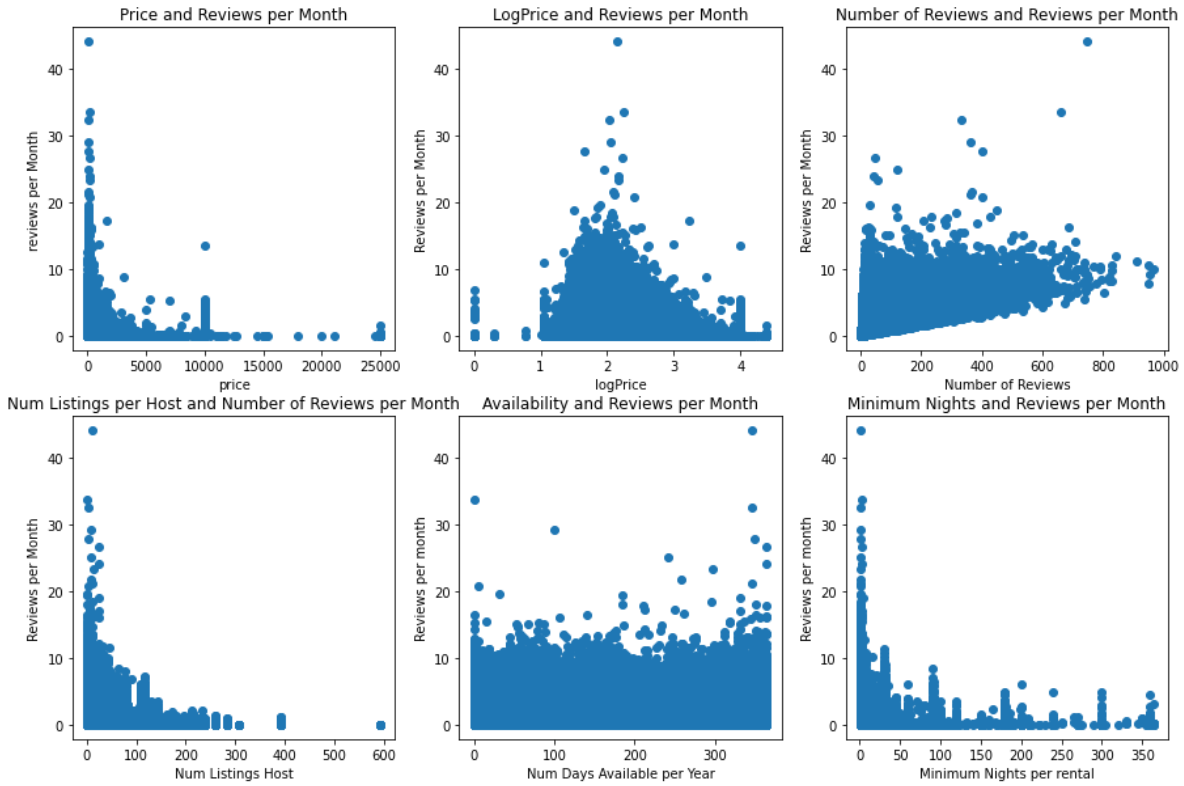
The Base Data

Column	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 neighborhood lies
Neighbourhood	Name of the neighborhood
Latitude	Latitude
Longitude	Longitude
Room_type	Room type (Entire home/apt, hotel, private room, shared room)
Price	Price per night
Minimum_nights	Minimum number of nights to book
Number_of_reviews	Total number of reviews of the listing
Last_Review	Date of last review
Reviews_per_month	Average reviews per month
Calculated_host_listings_count	Total number of listings by the host
Availability_365	Number of days the listing is available
City	Region of the listing

Distributions



Each Variable and Reviews per Month



The Data I Had to Add

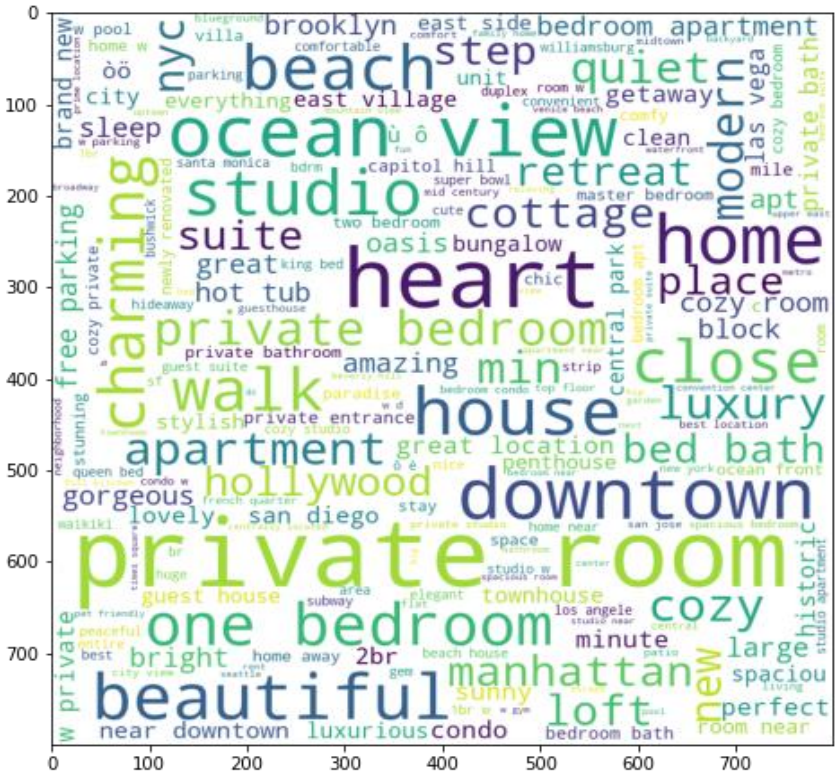
If there was any hope for a good model it would have to come from some basic language processing of the name column. So, I picked the most common words and said, “those sound like great features to add.”

They did the trick improving the R2 values by 0.1 on the test and train set.

Additionally, I added a few features that counted how many cafes, restaurants, and bars were nearby. This really didn’t do much, but I learned something valuable: I’m not a good real-estate agent and I have no idea what location features matter but I CAN get you location features.

To get this data and not have my computer explode in the process, I decided to use location centers from Kmeans clustering do each search in foursquare. Saved me several hypothetical pennies.

The Word Data

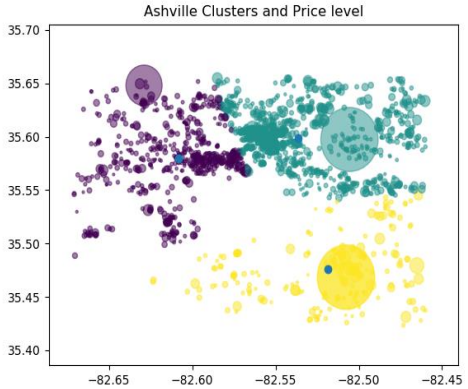


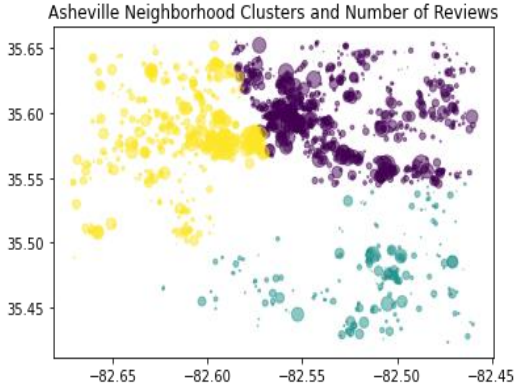
Focusing on Asheville

Why did I pick Asheville as the city to focus on?

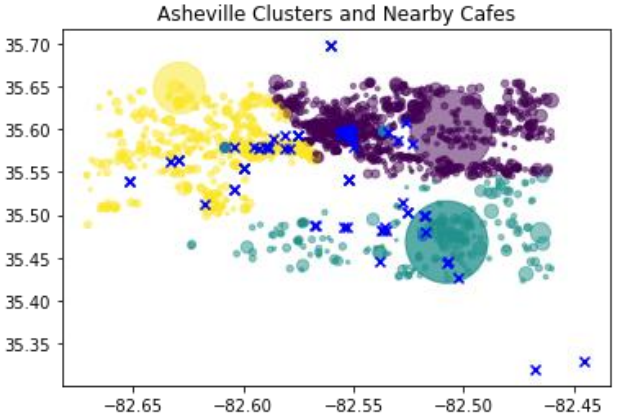
I didn't. The alphabet picked for me; however, the value in focusing on a particular city made getting location data simpler in terms of scale and my budget of exactly 39 dollars per month for this specialization. It also made visualizing the locations easier.

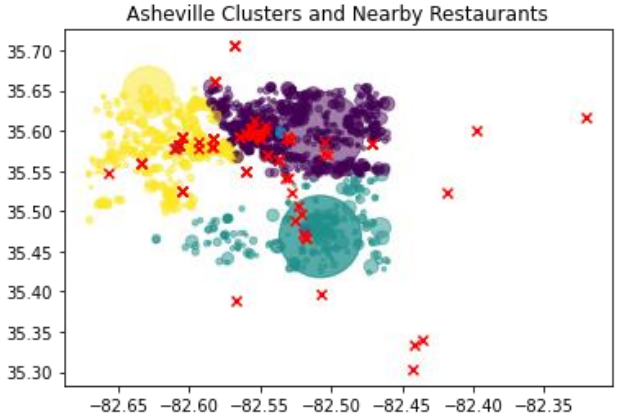
Here are a couple of scatterplots that show the location of each Airbnb and their neighborhood clusters. (Note the size of each circle in each plot represents price and reviews per month)

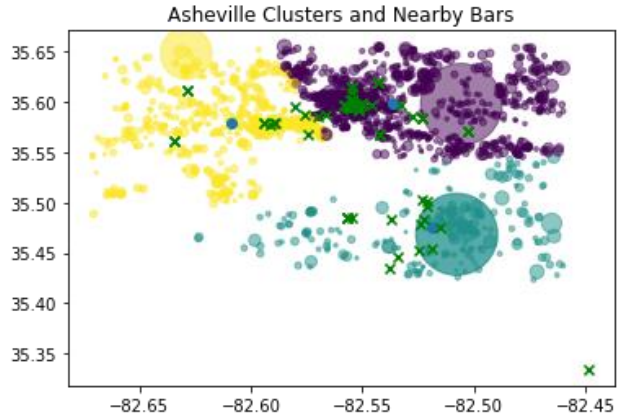




Locations of Local not so ‘hotspots’







Using this insight, we can get additional features by asking how many of each location type are nearby. Thanks Foursquare for helping me realize my wallet can be spared

What the models told us

A chart that shows R2 values for the training and testing set with and without location and word data:

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

Ridge and Lasso Models on Asheville

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

Conclusions

The regression model type nor hyperparameters seemed to indicate any significant differences in performance of the models meaning it comes down to features

If you’re a homeowner and don’t mind college students raiding your home for a day by all means, Airbnb is for you.

If you’re an investor I advise just being a landlord

If you still want to try

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

Link to Data

Dataset: https://www.kaggle.com/kritikseth/us-airbnb-open-data