Capstone 1 Milestone Report - AirBnB Price Prediction

Alexandra Michel

--Problem Statement--

AirBnb is a website that allows people to search and reserve short-term and apartment rentals, as well as become a host on the platform by opening up their own space to potential guests. All prices are set by the hosts, and guests/renters can narrow their search by different criteria to find a place to stay within their budget while still meeting their needs. As a host looking to earn supplemental income through AirBnB, it is important to know what the market price for your space is so that you can meet the demand of more consumers and ultimately have more guests book a stay with you.

There are third-party companies (Like this one, and this one) out there who provide AirBnB pricing for a fee; probably implementing similar techniques on data that is free to access like the data set I will use for this project. AirBnB also provides hosts with some advice when it comes to pricing, but who knows what attributes they are using to determine those prices, or how successful they are to attract guests (as hosts don't have to go with the suggested price, this would be hard to test!). AirBnB could be interested in this kind of exploration as a way to generate more revenue through their hosts, but more importantly, the average consumer with a spare room or living space is the ultimate consumer of the information this model would provide. Within the real estate industry, this would have a big impact for real estate agents and their clients (homebuyers and investors) looking to use a measure like this as part of the bigger picture in describing and analyzing the potential return on investment for a home purchase.

--Description of the dataset, how I obtained, cleaned, and wrangled it--

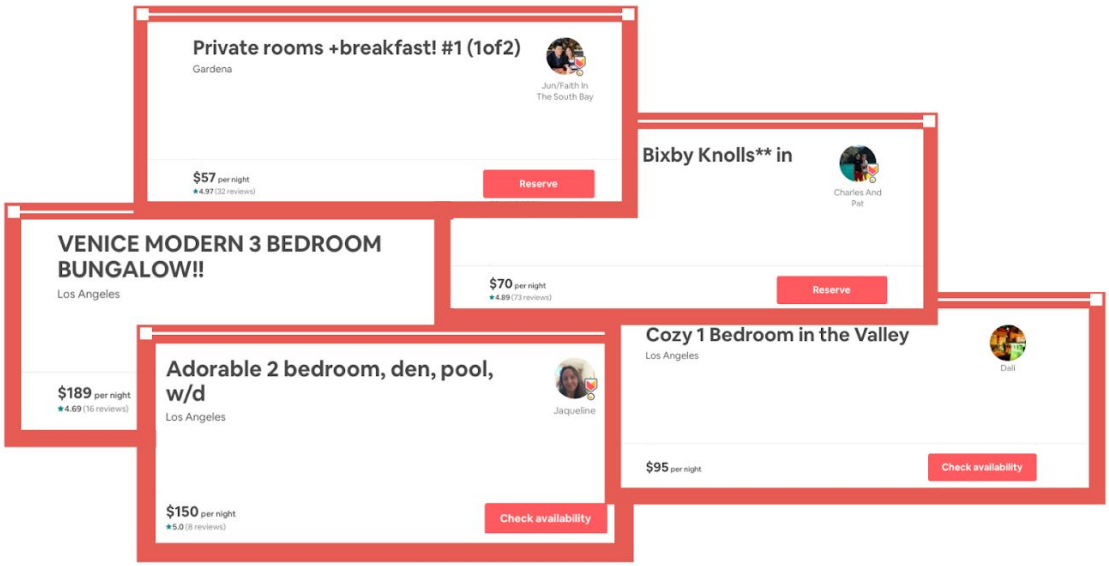
I obtained the data set from InsideAirBnB.com and loaded the .csv file for the Los Angeles, CA listings into a Pandas dataframe which I will call the "DF" for the sake of this report. Each row in DF is an available listing on the AirBnB site in Los Angeles, CA, and it has features related to listing criteria (bedrooms, bathrooms, beds, number of guests), as well as reviews, location, amenities, min/max number of nights. DF had many free-text features which I dropped for this initial attempt, but I would include them in the future as I learn to implement NLP techniques. I started out by also dropping a handful of columns which have mostly NaN values. There were also columns in DF that were closely related to each other (having to do with the total listings by

that same host) which I dropped and just left the `host_listings_count` column to account for that information. After these manipulations, I was already down from 106 columns to 60, reducing the dimension of DF significantly.

There were also sparse categories within `property_type` which I grouped together into `House`, `Apartment`, and `Other`. I was going to create a group called `Hotel` to include `Hotel`, `Boutique Hotel`, and `Bed and Breakfast`, but there would have only been 468 listings out of the total ~$38,000 listings so I decided it wasn't significant enough to keep.

The dollar-value columns `price`, `security_deposit`, `cleaning_fee` and the column `extra_people` needed to be changed from `string` to `int` values as well, stripping the dollar signs where needed and converting the `NaN` values to zeros.

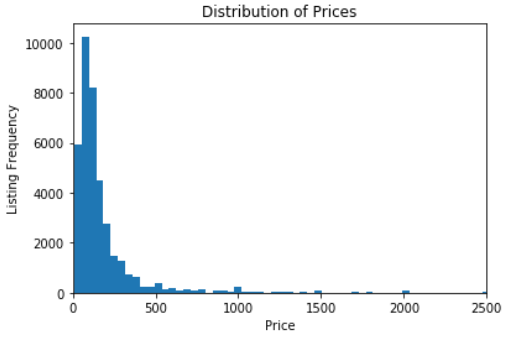
Once I had changed the `price` column to an `int`, I began exploring just the basic range of price values in DF. I came across a few listings with $0 for the nightly price which I thought was odd, and I also found that the highest nightly price was $22,000. I found the specific listings, and by using the listing URL was able to validate that, in fact, there are two luxury retreat listings at $22,000/night. However, the $0/night listings were mistakes. I know in general I would really just drop those entries, but because there were only 5 of them, I verified the prices posted as of 03-13-20 and manually inputted them into DF.



--Initial findings from exploratory analysis--

Question: What is the distribution of prices like?

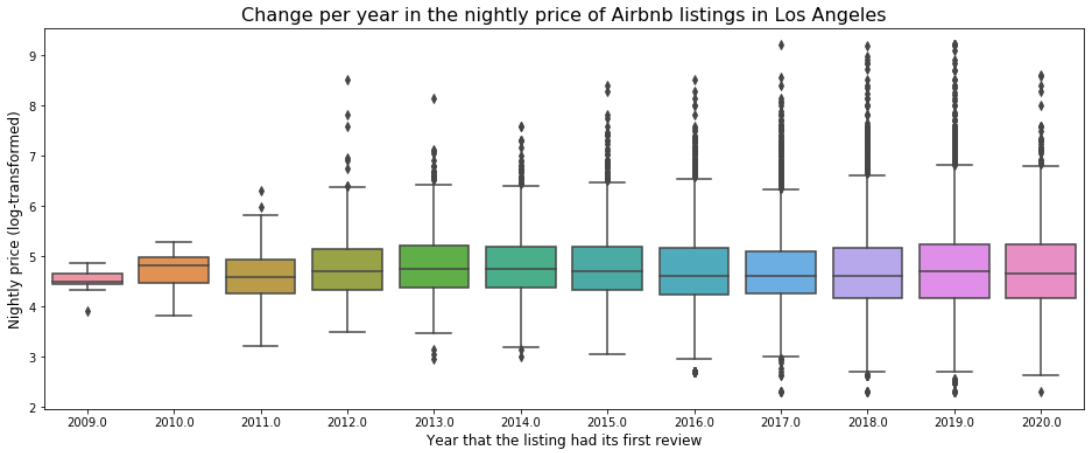
Because my model is working to predict the nightly price of an AirBnB listing, I decided to first take a look at what the distribution of prices looked like. The amount of listings above $2,500/night is less than 1% of the entire dataset, so I figured it would be best to inspect the distribution for listings below that price to give a better sense of the majority of prices.



From this graph and a calculation, we can see that 94.2% of the listings fall under $500/night. I also calculated that the median price is $110.0/night and the mean price is $226.88/night.

Question: How have prices changed over time?

The median price has only slightly increased over time as we can see on the graph below, because the maximum price has increased so much but there are still more listings at lower prices.

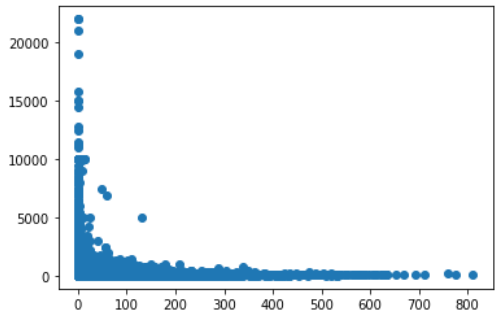


The maximum prices have increased dramatically from when this data started being compiled, it has caused the mean to increase from $93.11 per night to $180.77 per night, almost doubling in just 11 years of data. This begs the question: when did AirBnB start introducing the "Luxury Retreat" listings to the site? That could possibly explain some of this increase.

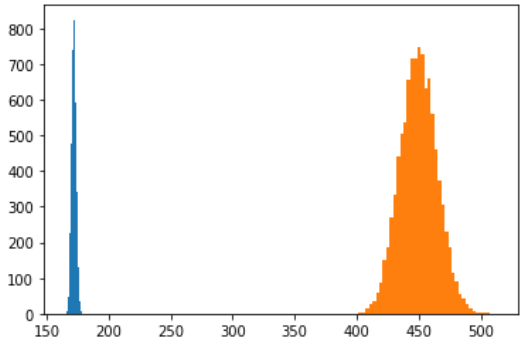
```
Mean nightly price of listings in each year on Airbnb in LA:
first_review
2009-01-01     93.11
2010-01-01    118.73
2011-01-01    122.68
2012-01-01    176.51
2013-01-01    176.62
2014-01-01    161.75
2015-01-01    164.66
2016-01-01    161.53
2017-01-01    163.79
2018-01-01    169.95
2019-01-01    183.05
2020-01-01    180.77
```

I also wonder if including the luxury listings or excluding them would lead to different results once I get to modeling this data. It seems like the higher prices could skew the data, but I wouldn't want to get rid of it prematurely.

Question: What is the distribution of reviews like? Does the absence of reviews affect price significantly?

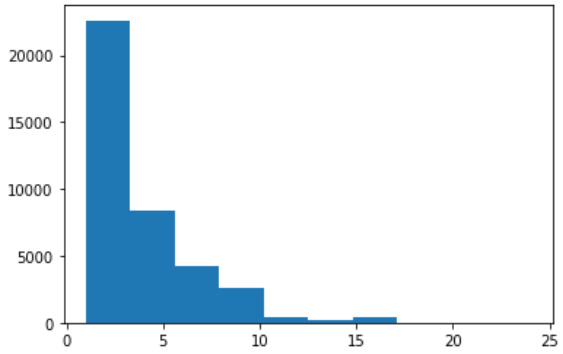


By separating the data into listings with/without reviews, it looked like listings with no reviews had a higher sample mean in general. We can also infer that a listing with no reviews is either a newer listing or a very expensive listing that not many guests have booked because of price. I decided to see if this is significant using bootstrapping:

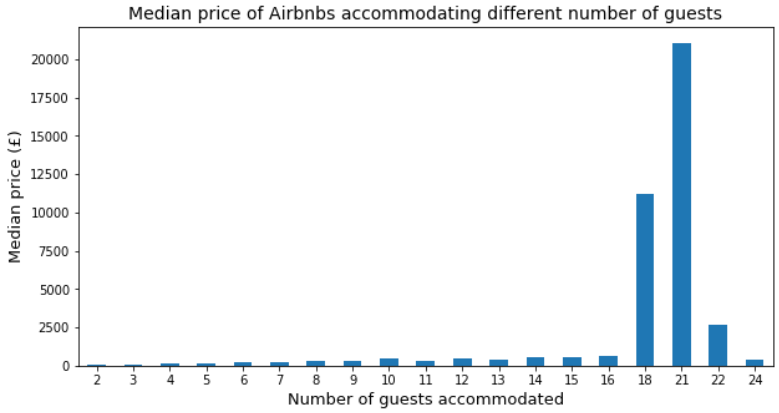


- 98% confidence interval for the sample mean for listings with reviews: [167.81, 176.41]
- 98% confidence interval for the sample mean for listings without reviews [416.57, 485.10]
- This tells us there is a 98% confidence that a listing with no reviews will be listed for $240 more

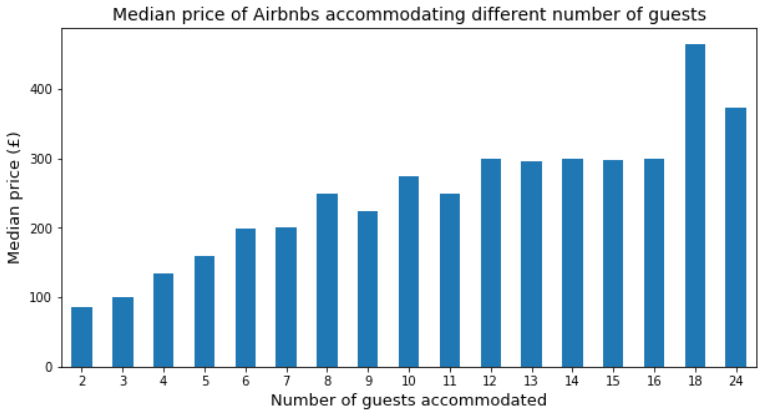
Question: What is the distribution of the number of people accommodated in an LA listing, and how does this affect price?



I first began by plotting the median price of listings with different accommodations, but quickly found that again the listings with higher prices were not allowing me to see what was really happening here

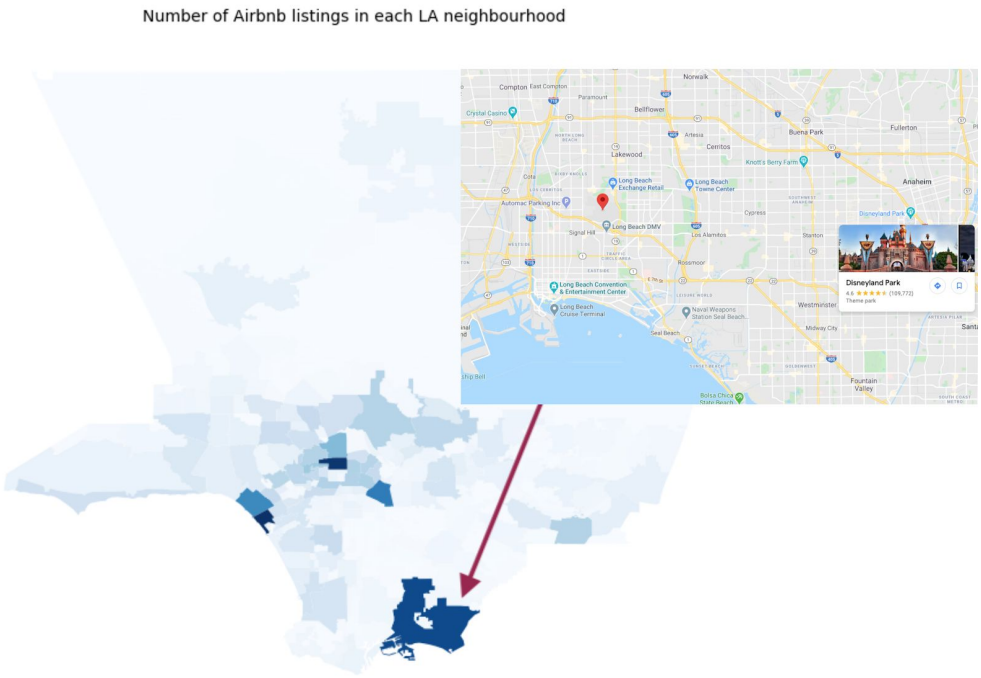


So I removed all listings above $500/night to see that we do have an increase, as expected, in price as the number of people the listing accommodates increases.

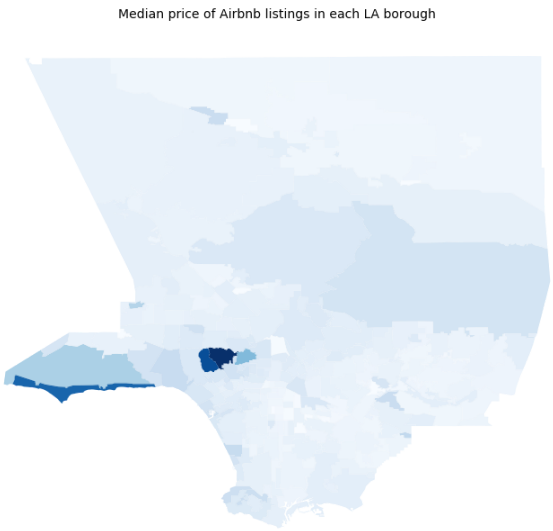


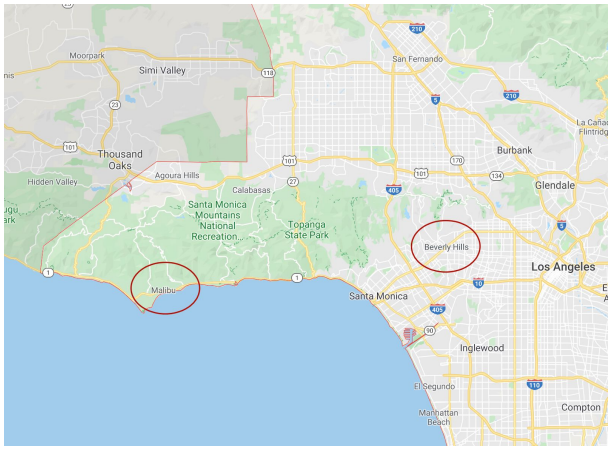
Question: Which areas of LA have the most listings, and which are the most expensive?

Before implementing the GeoPandas code to highlight these geographically, I made guesses that the most listings would be around Santa Monica or Hollywood, and near airports like LAX and Long Beach, and the most expensive would be the luxury listings in Malibu and Beverly Hills. I will say, I was pretty on target, but I forgot about one thing…



DISNEYLAND! One of the LA area's biggest tourist attractions. I don't know if this is why or not (Long Beach Airport is also in this area. Not as exciting.) but I would suspect the happiest place on earth has something to do with this. As for most expensive listings, yes right on point. We can see the Malibu coastline and Beverly Hills highlighted while the rest of LA is looking a bit pale...





--Next Steps--

When I got down to pre-processing I realized I had more columns I needed to deal with in order to take care of the one hot encoding correctly. I ended up separating the cancellation policies into 3 main categories, and review attributes into binned ratings ranges and NaN values. I also dropped the amenities feature because I couldn't figure out how to read in the large strings and encode the separate amenities, but this could be an area of further exploration to see if amenities help the model at all. For location I just kept the latitude/longitude coordinates and neighborhood name, and dropped the rest of the information for simplicity. I was going to keep the city/state to maybe incorporate multiple cities into the model, but simplified it for this approach.

My goal for this project is to be able to predict, with a good level of accuracy, the price of a listing based on the other meaningful attributes in the dataset. For the actual prediction model, it will be a prediction of prices, so I'll start out with a regression model and use feature selection to improve the model. This is not a classification problem, so I will proceed with using a random forest regression and possibly one other type, maybe Gradient Boosting, to create my model. So far I'm seeing that if neighborhoods don't play a huge role, I could see if dropping the luxury listings does anything to improve or worsen the accuracy of the model.

For further investigation it may also be interesting to do some unsupervised learning on neighborhoods and price (reverse to predict neighborhood based on price and other criteria).

--Supplemental Materials--

Github repository

Slide deck