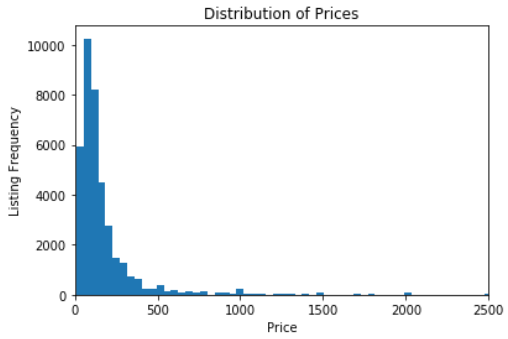
Capstone 1: Exploratory Data Analysis - AirBnB Price Prediction

Alexandra Michel

Question: What is the distribution of prices like?

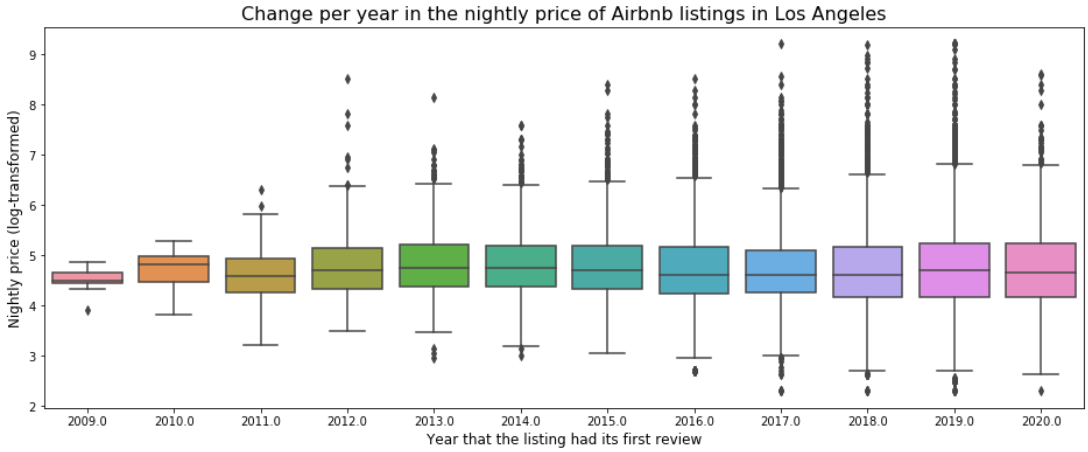
Because my model is working to predict the nightly price of an AirBnB listing, I decided to first take a look at what the distribution of prices looked like. The amount of listings above $2,500/night is less than 1% of the entire dataset, so I figured it would be best to inspect the distribution for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

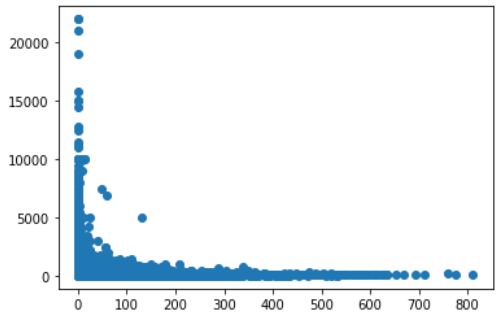


The maximum prices have increased dramatically from when this data started being compiled, it has caused the mean to increase from $93.11 per night to $180.77 per night, almost doubling in just 11 years of data. This begs the question: when did AirBnB start introducing the "Luxury Retreat" listings to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

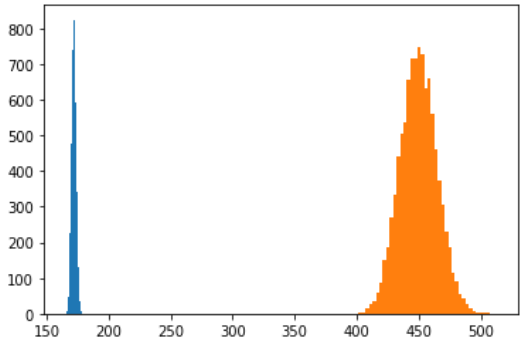
```
Mean nightly price of listings in each year on Airbnb in LA:
first_review
2009-01-01     93.11
2010-01-01    118.73
2011-01-01    122.68
2012-01-01    176.51
2013-01-01    176.62
2014-01-01    161.75
2015-01-01    164.66
2016-01-01    161.53
2017-01-01    163.79
2018-01-01    169.95
2019-01-01  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

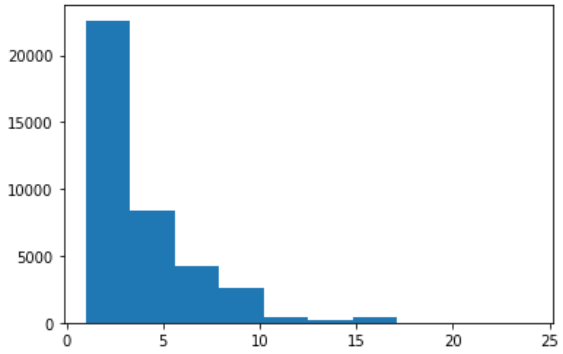


By separating the data into listings with/without reviews, it looked like listings with no reviews had a higher sample mean in general. We can also infer that a listing with no reviews is either a newer listing or a very expensive listing that not many guests have booked because of price. I decided to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

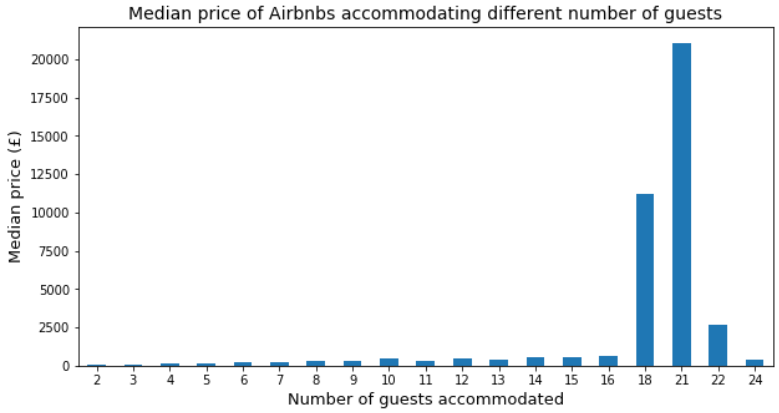


- 98% confidence interval for the sample mean for listings with reviews: [167.81, 176.41]
- 98% confidence interval for the sample mean for listings without reviews [416.57, 485.10]
- This tells us there is a 98% confidence that a listing with no reviews will be listed for $240 more

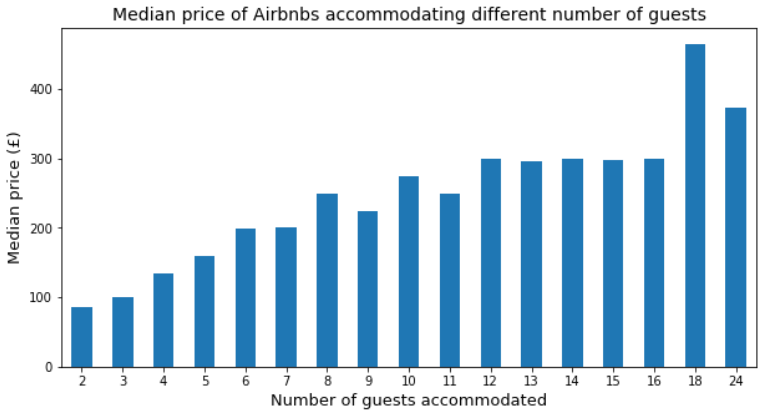
Question: What is the distribution of the number of people accommodated in an LA listing, and how does this affect price?



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

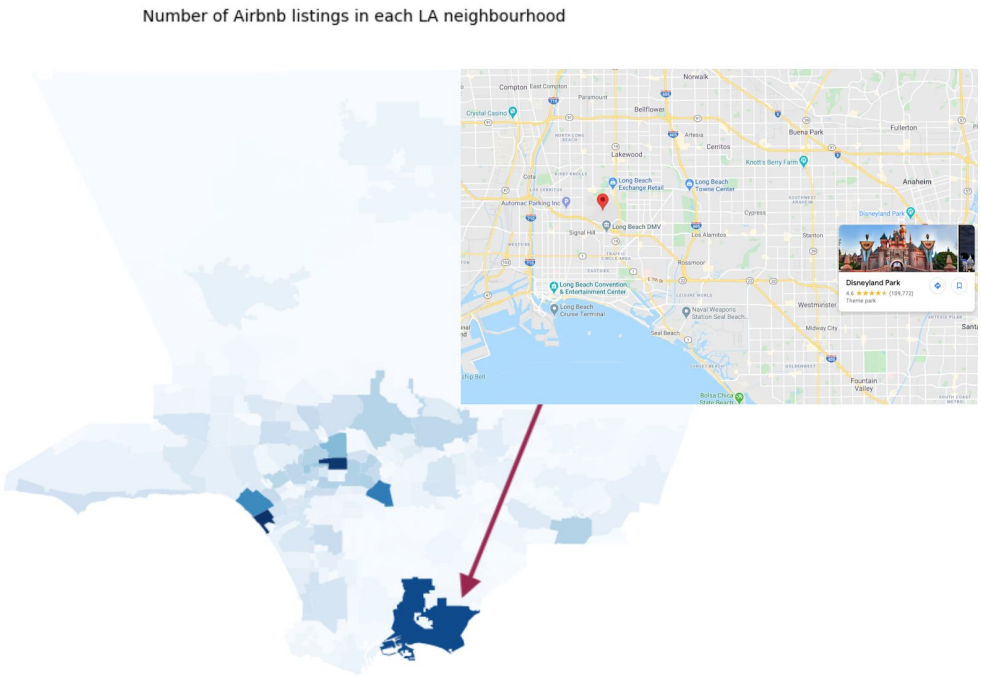


So I removed all listings above $500/night to see that we do have an increase, as expected, in price as the number of people the listing accommodates increases.



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...