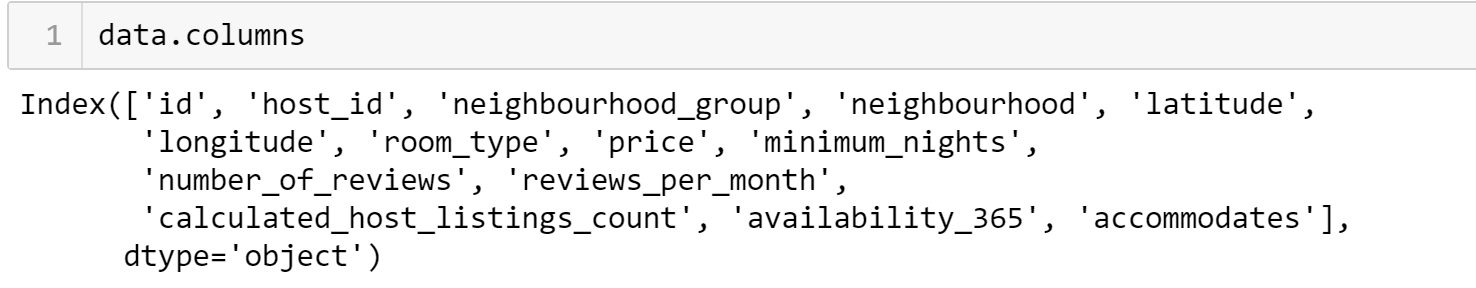
Capstone 3 Air B&B Report

By Thomas McMahon

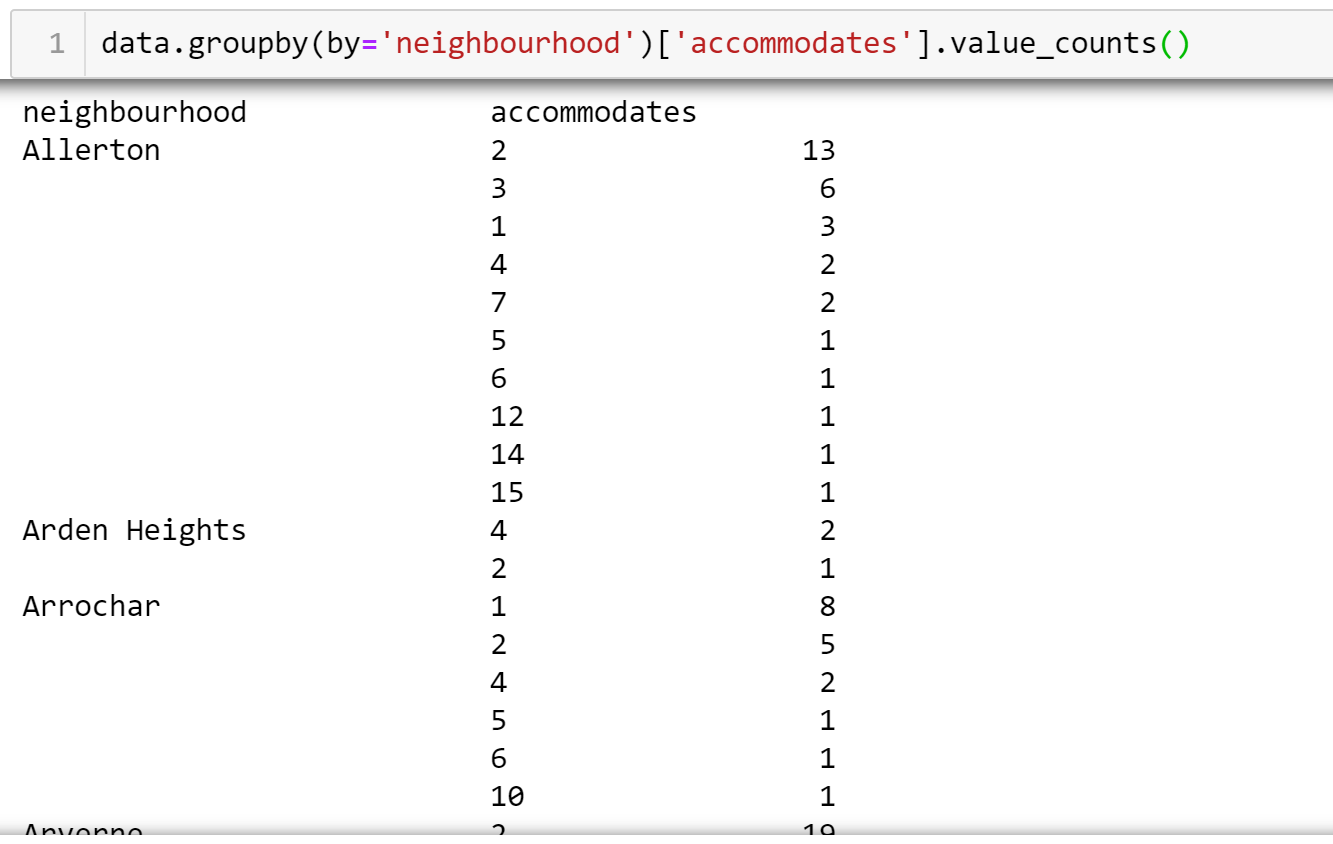
When I was growing up if you wanted to go on away on a vacation for a week or even just a couple days your options for where to stay was limited to staying with friends, staying at a hotel, staying at a motel, buying a timeshare. Now with the boom of the internet, smartphones, and other technology new bossiness sectors have been created. One if the sectors that has been created is short term publicly owned rentals. Now with apps such as Air B&B people are able to rent out any property, they own to anybody that they want. With these apps customers no longer have to rely on finding a hotel room and often can do it at a much cheaper rate, because you are renting from an individual and not a corporation. This allows people to travel the world without having to worry about every hotel room being booked for an event because there will always be people willing to rent you a room if the price is right. IN this project I will take a large city New York City and build a predictive model to predict how much the average price of an air b&b will be given how many people it can accommodate and where it is located. New York City is a perfect place to look because it has the five large boroughs as well as a very large dataset of rental histories. This project will be able to help both customers of air b&b and people looking to list their properties out to find a price range that will be able to complete in the open market while not being losing out on value by being under priced for the owners.

There are multiple public datasets that give all of the information that is needed to come up with the model. For the most part the data was clean, there were a lot of features to begin with, and a lot of them had thousands of missing values. When it came to cleaning the data the biggest issue was determining how to fill the null values, and more importantly if some of those features actually would help the model make predictions. A great example of this was the feature of last review. The columns had over two thousand missing values, the exact number matched the number of times that the column number of reviews was equal to zero. This was an easy one to fix filling all of the null values with zeros, but the question still stands does the time of the last review really matter when it comes to the listing price on an Air B&B. I decided it did not and ultimately removed the column to reduce noise in the model. I repeated this process with dozens of features until I had a dataset that was clean and the features were relevant to to the predictive model. 

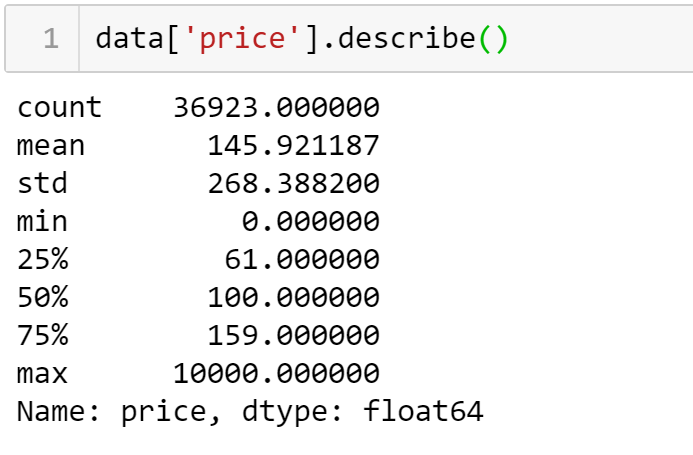
Once I had the features that I thought was relevant I started trying to find trends between the features. Some of what I was looking for was to see if any of the features could be broken down by other features, and to see if this information had any meaning. For example I wanted to look how many listings were there for each different number of people that a listing could accommodate in each borough. The purpose for this is to make sure that both features had enough data points across both of the features that the model would have enough to learn and predict vs having too little data points that could lead to the model using a feature that didn’t have enough points and caused an increase in bias towards that feature. Using this process I was able to find that multiple of the features that I thought were good features did not have the data that was needed and were dropped. After this process I was left with using the number of people that the listing could accommodate, the location of the listing as the borough and the specific neighborhood the listing is in and the price of the listing. It was not a lot of features, but it was the only features that I felt I needed to build an accurate model, while reducing potential noise from letting the model evaluate features that don’t have that much relevance to the price, although they could be used as a secondary if you wanted to be competitive in other ways, such as having all of the amenities that other listings have.

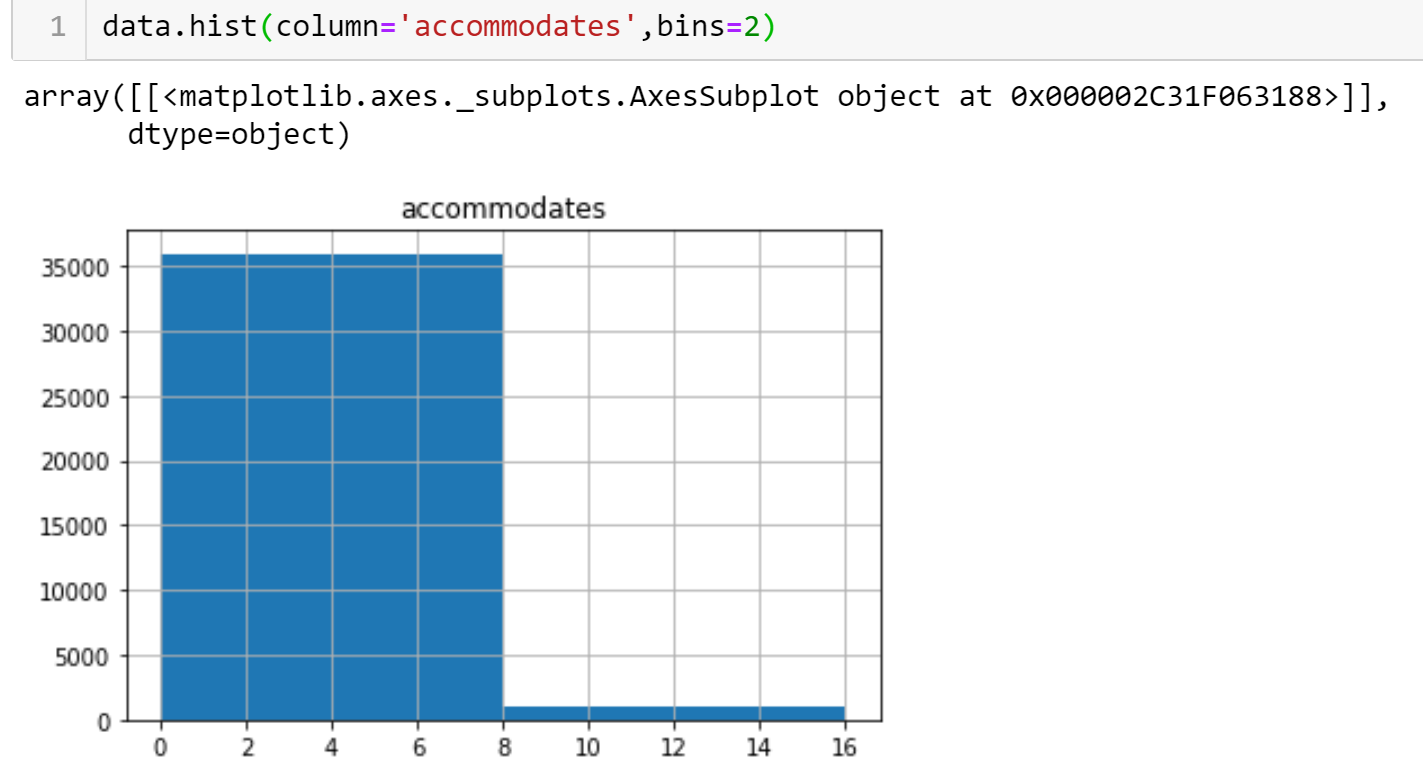
After looking at the trends and selecting the features that I felt were relevant it was time to start building some models. The results for the first set of models was very troubling. I used a linear regression as a baseline because I didn’t have that many features in the model that I felt that it wasn’t too complicated for a linear model. But when I ran the model the main score I was using to evaluate the model, mean absolute error, had a value of over one thousand. This meant that on average the prediction was off by over a thousand dollars. I though that the model might have been more complicated than I thought so I decided to run a random forest regression model. The results of that model were better, but not by nearly as much as I would have liked. At that point I new that I needed to go back to the drawing board and rework my feature engineering, and exploratory data analysis.

When I was going back through all of my features, I found what I though was the problem. The feature of the specific neighborhoods that the listings were located had a decent amount of total values, but not a lot of listings that accommodate the same number of people. This was what I believed was causing the issue by having the model use the specific neighborhood it caused there to be heavy bias to the little data that was available to be train on. Once I figured this out I removed the feature and remodeled the data.



Now with only the boroughs, number of people that the listing accommodates, and the price as the data points it was time to remodel the data. First when I used the linear model the results I got were much more promising. My mean absolute error dropped dramatically to just over $77. I still felt that this was too high so I I ran the random forest model again and again got only slightly better results. The random forest model gave me a mean absolute error of just under $75 dollars. I wanted to figure out why I still have a pretty high error so I went back and looked at the dataset, and the issue was that in the higher range there are less data points of listings that can accommodate more than 7 people and the prices of those listings were significantly higher than the listing price of the other listings. The max price of the listings was $10,000/ night with a standard deviation of $268. With this in mind I knew that models were doing a good job of making the predictions, but wanted to do more to se if I could make the model better.



What I wanted to see was if I separated the listings with the much higher than average amount of people able to be accommodated and focus on more similar listings if I could make the model better. To do this I needed to find the number of people that caused the huge shift. The way I did this was by plotting a histogram of the number of listings that could accommodate each different number of people. When I did this I found that the number where the drop off occurred was at 8. 

So from there I created two dataset one with all of the listings that accommodate 8 or more people and listings that were less than 8. When I modeled the two datasets I figured I was going to have an issue with the dataset with the larger accommodations. This is because it had less than 1000 datapoints, but none the less I wanted to still model the data. I got what I expected, which was a much larger mean absolute error of over $240, but that was to be expected with the limited data. On the other hand the dataset with the smaller accommodations saw a drop in mean absolute value to $62.73 witch is almost 20% better than the model with all of the datapoints.

In conclusion the models defiantly worked and can provide value to people either looking to rent out their property or people that and looking for a place to stay. I still believe that this model can get even better with increased data. The biggest issue that I had with the model was the fact that on the higher end I didn’t have enough data to accurately predict the price of a listing that accommodates a large group of people.