**Reflection Report**

IST 5520 - Fall 2021: Group 5

Members: Xinyue Chen, Nathan Dierkes, Joe Distler, Viraj Vilas Rane, Abby Ross,Yinkai Xiong

Our project was to analyze a dataset from AirBnB, a rental platform provider, and develop models that would be useful in their company or to their customers. The data was collected and provided by AirBnB, and posted publicly online for anyone to use. Our project scope was limited to one city in the United States: Chicago, Illinois.

At first look, the dataset has a large number of variables (74) which will cause a dimensionality issue later in our analysis. We were able to remove 13 of these variables by simply looking at the values and determining that they would not be useful in the analysis. For example, some columns were completely empty, others contained duplicate information, or url links. However, one item that we struggled with was reducing the dimensionality of the problem in order to obtain high-performing algorithms later on.

During the initial exploration and basic analysis of the data we found a number of interesting items within it. While a large number of the hosts (customers that used the platform to rent out their properties) only had 1-2 listings, there were some hosts that had as many as 69, 89 and even 260 listings. As you can guess these data points were outliers within this variable, but it is still interesting. Another thing we found interesting was that a large majority of the hosts had a very short response time. We expected to see a range of response rates within the data, but most of them responded within one hour of the request or message from the potential renters. Lastly in this initial exploration, we discovered that most of the properties were listed for $380 per night or less; however, there were multiple that were listed for well over $750 per night and maxed out at $1000 per night.

Once we felt that the basic analysis was complete, we moved on to comparing variables to each other in order to understand the general trends and continue to look for outliers in the data. During this stage, most of the trends we analyzed followed common sense logic as to how one would think they should. For example, our first comparison was the price of the listing compared to the number of amenities that were included in the ‘amenities’ field for the property. As one would expect, the price showed a positive relationship to the number of amenities. Similar trends were observed between the average rating compared to the response time; with a lower rating corresponding to longer response times. The average rating was positively correlated to the hosts’ response rate; i.e. the more a host responded to all messages and requests, the more likely it was they would have a higher average rating.

We then took the free text fields, ‘description’ [of the property], ‘neighborhood\_overview’, and ‘host\_about’, and performed a word cloud analysis on them. Word clouds are depictions of common words found in each of the fields, which provide a quick visual reference to the most commons words by making those larger compared to the others. Interestingly one of the more common words that showed up in the description field was the word “vaccinated.”

Our group wanted to try using more advanced graphs than the simple scatterplots and bar graphs, so we created joint plots to look more closely at other variable relationships. These graphs again showed the relationships that one would expect there to be between the variables. The average rating of a host was higher for the properties that were closer to the center of the city of Chicago. Additionally, the properties located closer to the center of the city on average were unable to accommodate more people.

One of the more interactive and interesting elements we were able to use was the heat map. Looking at raw data is one thing, but the heatmaps quickly let us understand the scope of the locations as well as how the price of the listings was correlated to the location on a map.

Our model building explored quite a number of different algorithm techniques, involving k-nearest neighbors, random forest ensembles, and decision trees. We used the k-nearest neighbors to build a model for both rating prediction and price prediction. Both of the models were tuned by adjusting the ‘k’ neighbors parameter. The models had fairly low performances, and we believe this to be from the large range of values in the rating and price variables. In addition, the dimensionality of the models was quite high. Next, random forests were used to predict the preferred property type (e.g. entire home, private room, etc.). This model was slightly higher performing with over 70% accuracy. Last, decision trees were used to predict customer group sizes as well as customer satisfaction and superhost prediction.