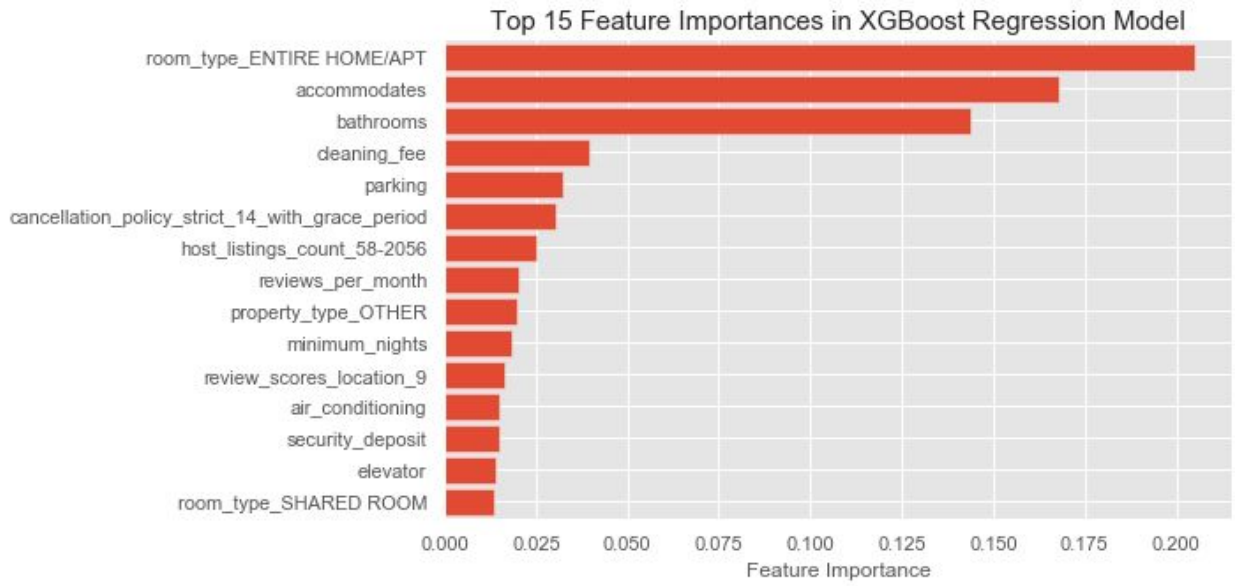
Capstone 1 - Statistical Analysis

Zach Palamara

Significant Features in the Data Set

After running a XGBoost regression model, I was able to determine which features in my dataset were most important in predicting the price of each listing.

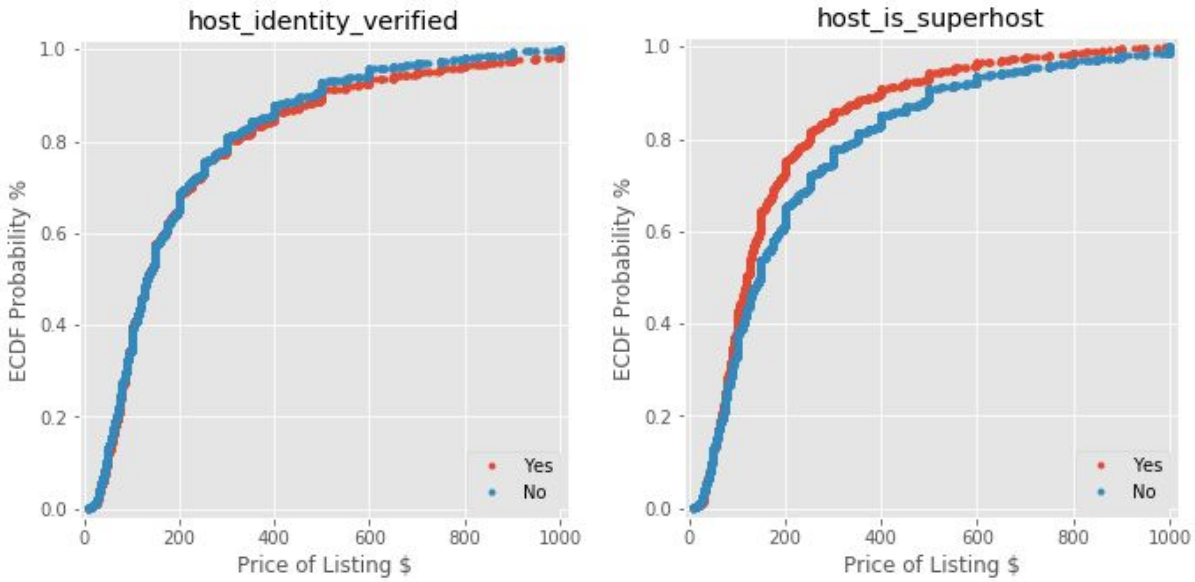


The top 3 features account for the majority of the influence on the model. The most important feature seems to be whether the listing is an entire house/apartment, which is not surprising considering that larger properties are more likely to accommodate more people compared to a shared room or something similar. Not surprisingly, the number of people that the listing accommodates is the 2nd most important feature followed by the number of bathrooms, which are features that have collinearity.

Subgroups of Data

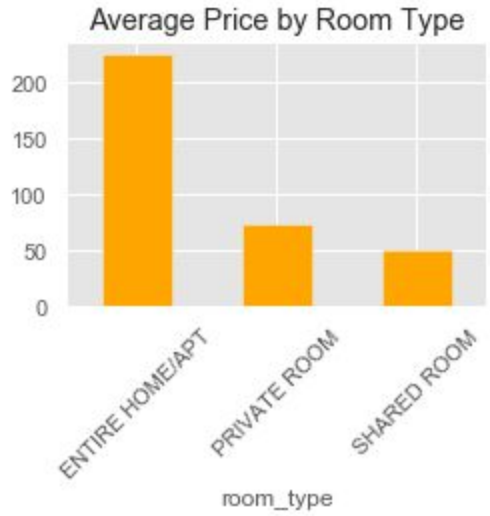
Superhosts

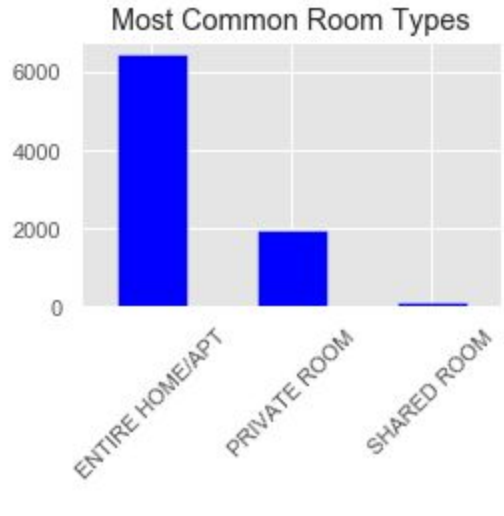
I wanted to explore whether or not being a superhost was valuable in terms of being able to charge a higher price for your listing. However, the ECDFs about the hosts indicate that prices are actually higher for non superhosts vs.superhosts.



Property/Room Type

It appears that there's a pretty big gap in average price between entire home/apartment listings vs. private rooms. Since this gap is so big, it might make the most sense to develop two independent models that are more representative of these two populations.

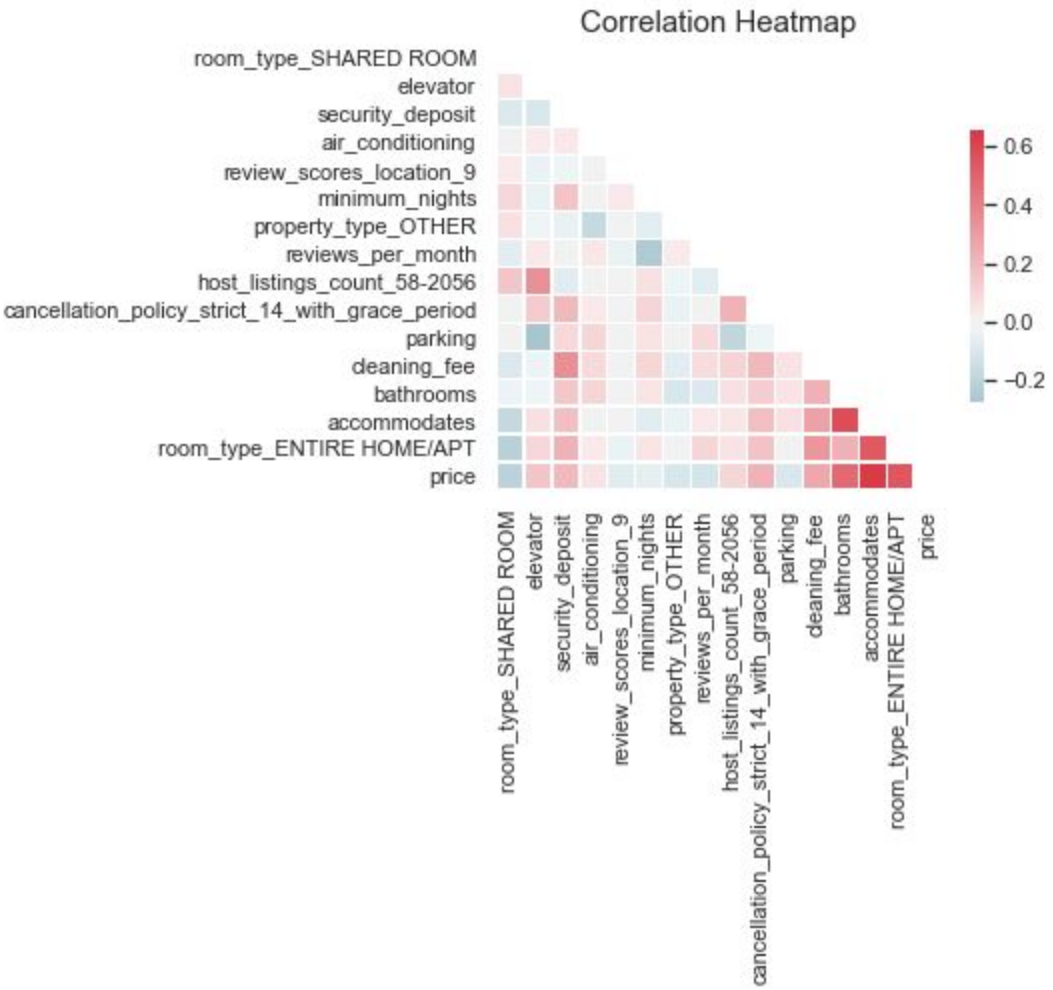

Average Price by Room Type
200
150
100
50
0
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
room_type


Most Common Room Types
6000
4000
2000
0
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM


Price Distribution
Entire Home/Apt.
Private Room
1000
800
600
400
200
0
Frequency
3
4
5
6
7
Log Transformed Price

Feature Correlation and Feature/Target Correlations

I was able to generate a correlation heatmap for the top 15 most important features using the seaborn library.



Positively Correlated Features

This heat map reaffirms the feature importance figure (above) generated in the XGBoost model, which shows that price is most strongly positively correlated with the accommodations, bathrooms and entire home/apt. features.

Negatively Correlated Features

One interesting thing that this heat map suggests is that the location score rating is actually slightly negatively correlated with price. Since the majority of the listings in this data set have room types of entire homes/apt it does make sense. Parking and elevators are also negatively

correlated features, which makes sense since most of the listings that have elevators are probably located closer to downtown where parking is always more scarce.