Data Interrogation and Predictive Analysis

Airbnb Prices

My statistical question was what drives an Airbnb price per night in New York. I also wanted to answer the hypothesis is the mean price per night was higher in Manhattan than it was in Brooklyn. I tested many models to predict price but, in the end, I found the best predictive variables to be neighborhood, room type, minimum nights, reviews per month, and availability. This model had the highest R-Squared and Adjusted R-Squared at .491 and .486 respectively. I also found that the mean price per night in Manhattan was higher than the mean price per night in Brooklyn. I found this out by doing a one-way hypothesis test comparing the mean price per night. The p-value of this hypothesis was less than .05 so the null hypothesis was rejected.

I think there could be more analyzed on with a log transformation. I like that the variables are easier to understand without doing a log transformation but for this data set, it may be needed. One variable that may have helped was average reviews. The data set included reviews per month but did not include what ranking was given.

I made an assumption to group all of the minimum nights stay together but the histogram was bi-modal. There was a peak at 1 night per stay and another at 30 nights per stay. It seemed like some Airbnb’s operated on a monthly basis so these would be priced differently.

One challenge I faced was understanding how to implement a log transformation into my analysis. I researched this topic but do not feel comfortable yet preforming it. Overall, I am proud of this project and fell like the analysis can be used to understands what drives price of Airbnbs in New York City.