**Name:** David Brehm

**Date:** 03/05/21

**Title:** Final Project Step 3 – NYC Airbnb Data

As someone who likes to travel, Airbnb has always been an interesting concept to me and something I’ve utilized extensively. It always seemed so much more appealing and immersive to stay in local housing instead of a hotel. Just like with any other type of housing, the prices of Airbnb’s can fluctuate. This dataset was appealing to me both for the geospatial aspect as well as a way to hopefully learn more about what impacts the price of a listing.

**Summarize the problem statement you addressed.**

With a dataset of 2019 New York City Airbnb listings, I wanted to model the prices and see what the best predictors of price were. Going back to the geospatial component, I also wanted to see if the prices aligned with neighborhood groups or another clustering method.

**Summarize how you addressed this problem statement (the data used and the methodology employed).**

The data was initially approached to observe missing values and outliers. After analyzing those, I removed variables that either wouldn’t contribute to price such as listing ID, or variables I couldn’t use. For example, the categorical variable neighborhood had too many values to model, so I dropped that and just used neighborhood group. From there I built a multiple linear regression model using all eight remaining variables. MSE, RMSE, and residuals were calculated, as well Cook’s distance. I went back and did a recursive feature elimination as another way to see which variables were important, as well as seeing the change in RMSE for different numbers of variables.

To build clusters, I started by plotting the Total Within-Clusters Sum of Squares and utilizing the elbow method to select the number of clusters for K-Means clustering. Choosing five seemed about right, and it conveniently matched the number of neighborhood groups in this dataset. After creating maps to display the K-Means clusters and neighborhood groups, I created another map for the prices.

**Summarize the interesting insights that your analysis provided.**

Based on the initial model and the Recursive Feature Elimination, room type was the strongest predictor of price. This seemed to also be supported by the plot of the average price by neighborhood group and room type, with “entire home/apt” clearly having the highest average price. The result makes sense,

I was a bit surprised that neighborhood group wasn’t a larger factor. The map and model showed that it had importance, but I originally thought it would have the largest impact on price.

**Summarize the implications to the consumer (target audience) of your analysis.**

Lowering your room type is the most effective way to lower your price. The number and frequency of reviews might be helpful for you to gain more context on a listing, but they aren’t affecting price as much. Location can have an impact in specific locations, but outside of those specific areas the effect drops.

**Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.**

A clear limitation on this data is a lack of review scores. When looking at reviews I definitely put more weight on the general score over the number of reviews or how recent they are. I’m not sure how much that would end up affecting price, but I imagine it would have a larger impact than the review statistics that were included here.