STAT 4355 Project Proposal

Team Name: The Superhosts
Armale Khan, Haeun Kim, Lakshmipriya Narayanan, Emily Painter

Source: https://www.kaggle.com/datasets/ybra1993/ab-ny-august-2019

Analysis Goal

The goal of our project is to use multiple linear regression to identify which features of Airbnb listings in New York City affect the price of a one night stay. By determining the typical price of an Airbnb listing using its features, potential hosts looking to list a new property could use our analysis to ensure they appropriately value their own property, and vacationers can ensure they are being charged a fair price. We will analyze and visualize this data by making plots using the programming language R.

Data Description

Our dataset details Airbnb listings from August of 2019 located in New York City. Airbnb stands for 'AIR Bed and Breakfast' and is an American company that lets owners rent out their property for tourists/travelers to stay. Our dataset has 106 variables and 48,864 instances. Given our dataset has such a large number of variables and that many of them are unusable for regression, we will only consider a subset of the features available in the dataset. If time allows and if our analysis suggests that more regressors would be of value, we may refer back to the original data and consider more regressors. Our response variable is price and our regressor variables are listed below.

Variables

- **Review_score_rating:** An aggregate rating out of 100 given by reviewers for the listing
- **Number_of_reviews:** Total number of reviews for the listing
- **Availability_365:** Total number of nights the listing was available in the last year
- **Minimum_nights:** The minimum number of nights you must reserve for the listing
- **Square feet:** The square footage of the listed property
- **Beds:** The number of beds available
- **Bedrooms:** The number of bedrooms (may not match the number of beds)
- **Bathrooms:** The number of bathrooms
- **Accommodates:** How many people can be accommodated overall
- **Host listings count:** Total number of listings per host
- **Host response rate:** Frequency the host responds to those who inquire about the listing
- **Host since:** The date the host first listed a property
- **Price (response):** Price in dollars for one night

In addition to these potential regressors, we will also use the **latitude**, **longitude**, **amenities,** and **host since** variables to create continuous variables that we can use in our regression (detailed in step 2 of our data analysis plan). We will also use the **zipcode** variable to create a subset of the data that is focused around central New York City.

Data Analysis Plan

1. *Data Cleaning:* Before performing any regression, we will remove any NA fields and drop the majority of the variables from the dataset that we will not be considering in our model.
2. *Creation of New Variables:* Four variables– **host since**, **amenities**, **latitude**, **longitude** –cannot be used in their original form. For the **host since** variable, we plan to create a new variable **num_years** for the number of years the host has had at least one property listed on Airbnb. For the **amenities** variable, we plan to create a new variable called **num_amenities** that tracks the total number of amenities detailed for each listing. For **latitude** and **longitude**, we plan to calculate a new variable **distance** which will record the distance of the listing from Central Park. This is how we will factor in the concept of location in our model of the full dataset.
3. *Modeling the Full Dataset vs a Subset of the Data:* In addition to considering different regressors in our models (detailed below), we also plan to compare our regression analysis of the total dataset to regression analysis of a subset of the data; specifically, we plan to analyze a subset of the data of only a select few zip codes that are all near central New York City. This subsetting of the data will eliminate the need for the **distance** variable as all instances in this subset of data will be comparable in location.
4. *Full Model Analysis and Dropping Individual Regressors*: After we have cleaned the data and created the relevant new variables, we will create a full model with all the potential regressors and examine the individual contribution of each regressor for both the full dataset and the subsetted data. Any regressors that appear to not be significantly contributing to the model at this stage will be removed.
5. *Reduced Model Analysis and Examining Subsets of Variables:* After we have conducted a full model analysis, we will create multiple reduced models to compare which subsets of variables are more impactful than others using ANOVA and determine if any further variables should be removed from the model.
6. *Model Comparison and Conclusions:* At this stage, we will compare the best models we have created with one another and determine which regressors most accurately model the response variable, price. From here we will be able to conclude what factors of those included in our dataset influence the price of an Airbnb listing in New York City (pre COVID-19). We will also compare our findings from the full dataset model with that of the subsetted dataset model and draw conclusions about the efficacy of our original **distance** variable based on whether or not the two models significantly differ.

Team Responsibilities

All members of our team plan to work on data cleaning and analysis together to go over all of the variables and determine any potential candidates for regressors. We will develop the original variables and the full model together, then we will delegate specific subsets of variables to each group member. We will reconvene with our individual analyses and determine the best models and draw conclusions from our individual analyses as a group.