**Final Project: New York Airbnb Prices**

**Objective**

We aim to find what specific variables impact the nightly price of an Airbnb in New

York in the year 2019.

**Data and Causality**

The dataset was pulled from Kaggle. Below details the causality of each of the thirteen variables:

* ID, Name, Host ID, Host Name: We deleted these, because those variables are meaningless for price per night as there are unique keys.
* Neighborhood Group: We believe it would impact the price of an Airbnb due to housing in a good location like downtown would have a higher price than the others elsewhere.
* Neighborhood, Latitude, Longitude: We did not utilize these factors because they are more granular location variables which we have already described by Neighborhood group.
* Room Type: Intuitively booking an entire house would cost more than booking a private room.
* Minimum Number of Nights:  The minimum number of nights was kept, and we assume the minimum number of nights booked would have a negative correlation with the price because the greater number of nights one booked is more possible to have a lower price per night.
* Number of Reviews and Last Review Date: Our assumption is that the number of reviews would have a positive effect on the price per night because more reviews imply higher popularity or at least the Airbnb being rented more often. The last review date only gives us the most recent date and cannot give us any important information that will affect the price per night.
* Availability: This variable is somewhat similar to minimum number of nights. We’re going to include it for the same reason as well. We had expected some multicollinearity between the two however the scatterplot of the two didn’t suggest that.

Thus, our final decision was to keep neighborhood group, room type, minimum nights, number of reviews, availability day of the whole year and price per night. Price per night is trivially our dependent variable in this, therefore the other six are the independent variables.

**Data Cleaning**

Although there weren’t any missing our data had some extreme variance. In the dependent variable there were many outliers. After researching online, we found out that the average price per night in NYC is $230 and the average travel day is 7 days. Since it is the average, we decided to set a range of price from $50 to $500 and a range of travel day from 1 day to 14 days. Because we want to measure what price a traveler is usually willing to pay for Airbnb per night on a short trip to New York. In addition, most of our dataset is in these two ranges. We also still have more than enough data, 37927, after we deleted data beyond these two ranges.

**Hypothesis Testing: Outliers in Number of Reviews**

Due to the overwhelming number of outliers in the variable number of reviews we deemed it fit to see if that led to a difference in price. Conducting a hypothesis test suggested there certainly was some sort of difference between the two, thus we will include this as an independent variable in our regression model.

**Multicollinearity, Interaction and Transformations**

After plotting with a scatterplot all the numeric variables, independent and dependent, it was visually evident that multicollinearity was absent. Apart from that, since we also have enough data points, multicollinearity shouldn’t be an issue. There was also no evidence that a transformation was needed for any of the independent variables when plotted against the dependent variables. One thing that marks an interesting note is that although some variables are numeric and can range from zero to infinity, it didn’t necessarily mean there continuous. For example, minimum number of nights are only positive integers; you cannot have four and half minimum nights. This made us reconsider how we handled them initially but ultimately decided we’d handle them the same way as if they were continuous.

As only business knowledge and logic can determine if there is an interaction term, we went to do some research on Google. After we have examined each independent variable individually, we do not believe that two of the variables should have different slopes in addition to different intercepts. Thus, we won’t consider any interactions in our regression.

**Regression**

All of our independent variables turned out to be significant except for the binary column representing outliers in number of reviews. Neighborhood was by far the most impactful in the variability of the price in our model, Manhattan being the most expensive and Staten Island the cheapest. Although via the F-test our model is significant, the R-squared is only about 32% suggesting that there are independent variables missing that describe the change in price. With the data that we currently have we cannot responsibly build a more explanatory model.

Recalling back to the insignificance of the column representing the outliers in number of reviews, this is certainly an interesting case. The hypothesis test showing there to be a significant difference, however when introduced to the model that is no longer the case. This can be explained by the fact that the hypothesis test only takes into account that specific variable and no others, so it may seem that there is a difference but in reality, there are other variables causing that difference that our hypothesis test is blind to.

**Residuals**

The residual does not necessarily conform with the assumptions we’d like them to. The sum of them are essentially zero, but when plotted it seems that they tend to hug around the zero line. It should be noted there are over 40,000 data points so it is somewhat difficult to eyeball the what exactly is going on. There is a slight skew to the right in the residual histogram. The QQ-plot however tells the same story though. There is clearly some sort of trend that is seen from the plot. However, we already concluded that our model is certainly missing something apart from the data, considering this these residuals are relatively okay.

**Statistical Tools**

* Linear Regression Analysis
  + Pros: Explains factor impact, Tests factor significance, Predictive capabilities
  + Cons: May not be significant, Needs to meet residual assumptions, May not have substantial data to build model such as missing variables, Linear approach may not be the best approach
* Scatterplots
  + Pros: Can show multicollinearity, Shows trends, Shows if transformation needed between independent variable and dependent variable
  + Cons: Cannot plot categorical data, Correlation is not causation
* QQ - Plot
  + Pros: Shows if residuals are normal
  + Cons: Attempting to manipulate regression to assume normality according plot could lead to fabricated results
* Histogram
  + Pros: Shows normality of residuals or any data if need be
  + Cons: Large amounts of data re needed to graphically represent normality
* Residuals vs. Order Index
  + Pros: Show heteroscedasticity of residuals
  + Cons: If data was ordered prior to regression the order is no longer randomly ordered so it could show some artificial trend
* Hypothesis Test? (For Outliers)
  + Pros: Tests for difference in dependent variable (i.e. do outliers make a significant different
  + Cons: The test’s scope is limited to one variable (In reality there may be something else driving the difference)