# **Dataset – New York City Airbnb Open Data**

For those who don’t know Airbnb – it offers lodging, primarily homestays and tourism experiences. Just like Uber, the company does not own any of the [real estate](https://en.wikipedia.org/wiki/Real_estate) nor does it host any events. It acts as a middleman between the customer and the owner providing the accommodation and Airbnb receives commission from each booking.

### **Problem Statement**

Problem statement is to explore Airbnb cost in different areas of New York. What are the variables influencing the price of Airbnb stay.

Variables in the dataset and their meaning:

|  |  |
| --- | --- |
| **id** | **listing ID** |
| **name** | **name of the listing** |
| **host\_id** | **host ID** |
| **host\_name** | **name of the host** |
| **neighbourhood\_group** | **location** |
| **neighbourhood** | **area** |
| **latitude** | **latitude coordinates** |
| **longitude** | **longitude coordinates** |
| **room\_type** | **listing space type** |
| **price** | **price in dollars** |
| **minimum\_nights** | **amount of nights minimum** |
| **number\_of\_reviews** | **number of reviews** |
| **last\_review** | **latest review** |
| **reviews\_per\_month** | **number of reviews per month** |
| **calculated\_host\_listings\_count** | **amount of listing per host** |
| **availability\_365** | **number of days when listing is available for booking** |

### **Questions**

The questions I would like to ask on the dataset are:

1. How the price is varying in the neighborhood
2. How the price is varying with room type
3. What is the relationship between location and price
4. Are reviews influencing the price

### **Solution Approach**

I initially did the data cleansing steps like data type conversions, drop unwanted variables, duplicate data check and missing values check.

Later, I explored the categorical variables like room type, neighborhood and explored distribution of the variables using histograms.

I checked the correlation of price variable with other variables in the dataset. Correlation tells us the strength of the relationship between the variables.

Target variable with LOG transformation gave me normal distribution, so that’s appropriate transformation. I further explored the relationship visually using scatter plots.

Finally, I used linear regression to build a model to predict Price using other related variables in the dataset and checked for p-value and R-squared values. Second regression model with Log transformation gave me better model compared to previous model. So, I concluded my model with interpretation there.

### **Interesting insights**

What I found interesting was distribution of Price variable without transformation was skewed and with Log Transformation it appeared like normal distribution.

Another interesting insight is, Minimum number of nights stay is not impacting the Price of the hotel.

### **Limitations**

Though I concluded the final model with Log transformation on target variable, the final model didn’t explain the variance in the dataset completely. It could only explain the variance around 50%, so the predicted price values would be off by a certain amount. Our dataset also has outliers, what if the outliers are genuine values and the model couldn’t fit them completely.

### **References**

Kaggle - New York City Airbnb Open Data

<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

You can look at my code here as well:

<https://github.com/satsmanthani/DSC530/blob/master/DSC520-FinalProject%5B2%5D.Rmd>