Price Prediction Analysis of New York City's Airbnb Data

Introduction

Airbnb has seen a meteoric growth since its inception in 2008 with the number of rentals listed on its website growing exponentially each year. Airbnb has successfully disrupted the traditional hospitality industry as more and more travelers, not just the ones who are looking for a bang for their buck but also, business travelers' resort to Airbnb as their premier accommodation provider. New York City has been one of the hottest markets for Airbnb, with over 52,000 listings as of November 2018. This means there are over 40 homes being rented out per square km. in NYC on Airbnb! One can perhaps attribute the success of Airbnb in NYC to the high rates charged by the hotels, which are primarily driven by the exorbitant rental prices in the city.

Data Wrangling technique steps:

1. What kind of cleaning steps did you perform?
2. How did you deal with missing values, if any?
3. Were there outliers, and how did you handle them?

Steps: Cleaning Datasets

After looking at the info() method of the dataset, I saw some NaN values, therefore need to examine missing values further before continuing with analysis.

Find out the which columns have null values. Hence using the "**isnull ().sum**()" function will show us how many missing values are found in each column in dataset.

a. In our case, missing data that is observed does not need too much special treatment. Looking into the nature of our dataset we can state further things: columns "**name**" and "**host_name**" are irrelevant and insignificant to our data analysis for pricing prediction. Also **"last_review"** and **"review_per_month"** need very simple handling.

b. Filling Missing Value **:** divide columns with missing value into several categories (based on the data to be filled in each of them).To elaborate, "**last_review**" is **date**; if there were no reviews for the listing - date simply will not exist. In our case, this column is irrelevant and insignificant therefore appending those values to '0'.

For **"review_per_month"** column we can simply append it with 0.0 for missing values; we can see that in "**number_of_review"** that column will have a 0, therefore following this logic with 0 total reviews there will be 0.0 rate of reviews per month. Therefore, let's proceed with removing columns that are not important and handling of missing data.

c. Dropping the insignificant or irrelevant columns from datasets. Using dropping **.drop()** method **"name"** and **"host_name"** not only because it is insignificant but also for ethical reason. Also, these names are unimportant to our data analysis.

d. Outliers : We have applied the IQR technique to know the how far the datasets is distant from each other's. "Outliers is a datapoint which is distant from all other observation" Using the Inter quantile range (IQR=Q3-Q1) technique to identify the outliers in describe method. We have dropped the text /categorical variable columns and ensure all columns are float64 format.

Note : Scatterplot , Boxplot , Z-score are the other method for outliers.

References:

- Clean and Tidy paper written by Hadley Wickham's.