Analysis of New York City's Airbnb Data

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?

Procedure for Data Wrangling technique

1. **Cleaning Steps:** 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, columns **"last_review"** and **"review_per_month"** need very simple handling.

   a.Using the dropping **.drop()** method "**name"** and **"host_name**" not only because it is insignificant but also for ethical reasons. Also, these names are unimportant to our data analysis.

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

2. **Filing Missing Value** : Divide the columns with missing value into several categories (based on the data to be filled with each other). To elaborate, **"last_review**" is a 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 is not needed. For the **"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 missing data.

3. **Outlier** : We have applied the IQR technique to know how far the datasets are distinct from each other.

   "Outlier is a datapoint which is distinct from all other observation"

   Using the Interquartile 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** : Scatter , Box , Z-score are the other method for outliers.

4. References:

   Clean and Tidy paper written by Hadley Wickham's.