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: After looking at the **info method** of the dataset, I saw some **missing** values in the column: **name, host_name,last_review,review_per_month**.

- Using the "**isnull ().sum**()" function will show us how many missing values are found in each column in the dataset.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 because it is insignificant but also for ethical reasons. Also, these names are unimportant to our data analysis.
- Columns **"last_review"** and **"review_per_month"** need very simple handling.

2.Filing Missing Value : Divide the columns with missing value into several categories (based on the data to be filled with each other).

- **"review_per_month"** and "**number_of_review"** column we can simply append it with 0.0 for missing values; **"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 to 0 .
- "**name**" and "**host_name**" filed with "NA " because these columns are irrelevant and not important to our analysis. There may be several reasons not to capture the name and host_name.Therefore, let's proceed with removing columns that are not important and handling missing data.
- Dropping the column "**id**" and "**host_id**" because it does not need too much special treatment. Again use info() method to check the each column counts and verify the missing values are treated correctly.

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 the described method. We have dropped the text /categorical variable columns and ensure all columns are float64 format.
- **Identifying the Outlier using** Scatter , Box , Z-score are the other methods for outliers.
- **Identifying Outliers with Skewness** : Ideally, the skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values.

- **Outlier Treatment** : In this technique, we will do flooring for lower(10%) values and capping for higher(90%) values. These values will be based on quantile-based **flooring and capping methods.** After i.e we have set np.where() method for price,number_of_review,reviews_per_month,calculated_host_listings_count,availability_365 and create new skewness columns.
- Later, we verify the outlier detection counts based on Z-Score.

Note : References: Clean and Tidy paper written by Hadley Wickham's.