**Introduction:**

Airbnb is an online-based marketing company that connects people looking for accommodation (Airbnb guests) to people looking to rent their properties (Airbnb hosts) on a short-term or long-term basis. The rentals properties include apartments (dominant), homes, boats, and whole lot more. Since its inception in 2008, Airbnb has steadily risen in terms of revenue growth and its range of service provisions. As of 2019, there 150 million users of Airbnb services in 191 countries, making it a major disruptor of the traditional hospitality industry (this is akin to how Uber and other emerging transportation services have disrupted the traditional intra-city transportation services).

Airbnb generates revenue by charging its guests and hosts fees for arranging stays: hosts are charged 3% of the value of the booking, while guests are charged 6%-12% per the nature of the booking. As a rental ecosystem, Airbnb generates tons of data including but not limited to: density of rentals across regions (cities and neighborhoods), price variations across rentals, host-guest interactions in the form of reviews, and so forth.

New York city (NYC) has an extremely active Airbnb market with more than 48,000 listings as of August in the 2019 calendar year (this corresponds to a rental density of 48000 rentals per 468 square miles, which equates to 102 rentals per square mile). This project focuses on the gleaning patterns and other relevant information about Airbnb listings in NYC. To be more specifically, the goals of this projects are to answer questions such as: (i) how are rental properties distributed across the neighborhoods of NYC (there are 221 neighborhoods); (ii) how do prices vary with respect neighborhoods, rental property types and rental amenities; (iii) more importantly, how well can machine learning models be trained to predict Airbnb rental prices using features such as rental property type, the  number of people a rental can accommodate, the number of available beds, and so forth.

**Context**

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

**Content**

This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

**Inspiration**

* What can we learn about different hosts and areas?
* What can we learn from predictions? (ex: locations, prices, reviews, etc)
* Which hosts are the busiest and why?
* Is there any noticeable difference of traffic among different areas and what could be the reason for it?

**Attribute information:**

* This dataset has around 48,895 observations in it with 16 columns and it is a mix between categorical and numeric values.
* These 16 columns provide a very rich amount of information for deep data exploration we can do on this dataset. We do already see some missing values, which will require cleaning and handling of NaN values. Later, we may need to continue with mapping certain values to ones and zeros for predictive analytics.
* 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.
* 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 is not needed.
* 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.
* Interesting, we can see that there is a good distribution between top 10 hosts with the most listings. First host has more than 300+ listings.

|  |  |
| --- | --- |
| **Features** | **Description** |
| id | listing id |
| host\_id | host id |
| neighbourhood\_group- | location |
| name- | name of the listing |
| host\_name - | name of the host |
| 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 |