**Airbnb Technical Doumentation**

Airbnb is a $75 Billion online marketplace for renting out homes/villas/ private rooms. The website charges a commission (3 to 20 percent, ) for every booking. Even though the prospects are sound, but there are critics who argue that this has driven up rent, and caused damage to the local communities living in the vicinity.

The data used in this analysis is the outcome of the quest to answer the question – How is Airbnb affecting the neighbourhoods? [Insideairbnb](http://insideairbnb.com/about.html" \t "https://www.analyticsvidhya.com/blog/2021/10/end-to-end-predictive-analysis-on-airbnb-listings-data/_blank)is an activist project, which has curated this dataset, to measure the impact of rentals housing on neighbourhoods and communities.

A lot of articles focus more on the machine learning algorithm/model, data cleaning, feature extraction and fail to explain the purpose of the model. Understanding the business model can help identify challenges that can be solved using analytics and data science.

Airbnb provides various rental options for different customer segments. Based on customer budget, they can either opt for an entire house or just a room or even better share a room. With a range of prices as low as 700 to as high as 50,000, comes a range of amenities, such as selection on a number of beds, bedrooms, kitchen, air conditioning, heating washing machine, breakfast, beachfront, gym, pool etc to name a few.

There are four major types of places:

1. Entire place
2. Private room
3. Hotel room
4. Shared room

For the hosts, Airbnb has a [superhost](https://www.airbnb.co.in/d/superhost" \t "https://www.analyticsvidhya.com/blog/2021/10/end-to-end-predictive-analysis-on-airbnb-listings-data/_blank)program providing exclusive benefits and higher visibility.

Hosts also have to comply with certain state/municipality rules and regulations to be listed on Airbnb. These rules are promulgated keeping in mind the safety of the residents living in the residential blocks, to prevent malpractices by individuals or a group. Airbnb has to comply with these rules as well to be able to operate in that city.

**How are rental prices decided?**

1. Airbnb service fee: Guest service fee charged by Airbnb—this helps Airbnb run smoothly and offer 24/7 community support
2. Cleaning fee: Fee charged by some hosts to cover the cost of cleaning their space (applicable to all countries except China)
3. Extra guest fee: Fee charged by some hosts for each guest beyond a set number
4. Security deposit: Certain reservations may independently require a security deposit requested by the host or Airbnb
5. Value Added Tax (VAT, JCT, and GST): Tax charged to guests who live in certain countries
6. Local taxes: Taxes charged based on the location of the host’s listing

## Exploratory Data Analysis and Predictive Modelling

**Data Cleaning/Wrangling:**

As always the most important part of an analysis is the data, as this data is scraped, the chances of missing values, odd HTML characters, sparse columns, high cardinal columns are high. For the purpose of this analysis, the focus is on 3 datasets only, Calendar, Listings and Reviews.

**Importing Data:**

Importing using pandas read\_csv. Calendar, listing, reviews are imported into df\_calender, df\_listings, df\_reviews respectively.

**Pandas Profiling:**

Writing code to print descriptive statistics, find missing values/ duplicates, memory usage, correlation is made easy using this [library](https://github.com/pandas-profiling/pandas-profiling). [Follow](https://www.analyticsvidhya.com/blog/2021/06/generate-reports-using-pandas-profiling-deploy-using-streamlit/)this article on Analytics Vidhya to know more about this awesome library. The profile provides a report on the below parameters:

1. Type inference: detect the types of columns in a dataframe.
2. Essentials: type, unique values, missing values
3. Quantile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range
4. Descriptive statistics like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
5. Most frequent values
6. Histogram
7. Correlations highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
8. Missing values matrix, count, heatmap, and dendrogram of missing values
9. Text analysis learns about categories (Uppercase, Space), scripts (Latin, Cyrillic), and blocks (ASCII) of text data.
10. File and Image analysis extract file sizes, creation dates, and dimensions and scan for truncated images or those containing EXIF information.

Install pandas profiling

!pip install pandas-profiling

Generate profile report :

The report can be saved as an HTML as well

## **Practice – Communicate Results and Findings:**

The findings from the EDA can be informative to either hosts or tourists or both. Write down a few findings from the EDA so that hosts can gain better ratings or higher prices. For tourists, findings can be around the best time to visit, or easy commute options, etc.

Another effective way to communicate results is to visualize data, histograms to visualize the spread, bar graphs to compare different neighbourhoods, and room types, line graphs to visualize price trends, seasonality, weekday weekend effects. Maps(Heat maps) can be used to highlight areas based on different attributes such as price, easiness to commute, bubble charts can provide availability information, etc. Visualizations provide an effective way to connect with the audience by the virtue of being appealing, vibrant, and easy to remember.