# Airbnb Case Study

## Problem background

Suppose that you are working as a data analyst at Airbnb. For the past few months, Airbnb has seen a major decline in revenue. Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.

The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset so as to increase the revenue

## End Objective

To prepare for the next best steps that Airbnb needs to take as a business, you have been asked to analyse a dataset consisting of various Airbnb listings in New York. Based on this analysis, you need to give two presentations to the following groups.

**Presentation - I**

- Data Analysis Managers: These people manage the data analysts directly for processes and their technical expertise is basic.
- Lead Data Analyst: The lead data analyst looks after the entire team of data and business analysts and is technically sound.

**Presentation - II**

- Head of Acquisitions and Operations, NYC: This head looks after all the property and host acquisitions and operations. Acquisition of the best properties, price negotiation, and negotiating the services the properties offer falls under the purview of this role.
- Head of User Experience, NYC: The head of user experience looks after the customer preferences and also handles the properties listed on the website and the Airbnb app. Basically, the head of user experience tries to optimise the order of property listing in certain neighbourhoods and cities in order to get every property the optimal amount of traction

# Methodology

Data Cleaning & Preparation for visualization and analysis purpose
Tool Used - Python

```
airbnb.head()
```

| | id | name | host_id | host_name | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_nights | number_of_revie |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2539 | Clean & quiet apt home by the park | 2787 | John | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 | 1 | |
| 1 | 2595 | Skylit Midtown Castle | 2845 | Jennifer | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 | 1 | |
| 2 | 3647 | THE VILLAGE OF HARLEM....NEW YORK ! | 4632 | Elisabeth | Manhattan | Harlem | 40.80902 | -73.94190 | Private room | 150 | 3 | |
| 3 | 3831 | Cozy Entire Floor of Brownstone | 4869 | LisaRoxanne | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | Entire home/apt | 89 | 1 | 2 |
| 4 | 5022 | Entire Apt: Spacious Studio/Loft by central park | 7192 | Laura | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | 80 | 10 | |

```
airbnb.shape
```

```
(48895, 16)
```

## Remove columns not significant in Analysis and check info

```
#remove name and id column as not used for analysis
airbnb.drop(["id","name"],axis=1,inplace=True)
```

```
airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   host_id                         48895 non-null  int64  
 1   host_name                       48874 non-null  object 
 2   neighbourhood_group             48895 non-null  object 
 3   neighbourhood                   48895 non-null  object 
 4   latitude                        48895 non-null  float64
 5   longitude                       48895 non-null  float64
 6   room_type                       48895 non-null  object 
 7   price                           48895 non-null  int64  
 8   minimum_nights                  48895 non-null  int64  
 9   number_of_reviews               48895 non-null  int64  
 10  last_review                     38843 non-null  object 
 11  reviews_per_month               38843 non-null  float64
 12  calculated_host_listings_count  48895 non-null  int64  
 13  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(6), object(5)
memory usage: 5.2+ MB
```

## Check Null values

```
airbnb.isnull().mean()*100
```

```
host_id                           0.000000
host_name                         0.042949
neighbourhood_group               0.000000
neighbourhood                     0.000000
latitude                          0.000000
longitude                         0.000000
room_type                         0.000000
price                             0.000000
minimum_nights                    0.000000
number_of_reviews                 0.000000
last_review                      20.558339
reviews_per_month                20.558339
calculated_host_listings_count    0.000000
availability_365                  0.000000
dtype: float64
```

- Host name we found that its missing by chance and need to get information from team ,but for now we kept it as it is
- We found that 'last_review' and 'reviews_per_month' columns are not randomly missing; rather, they indicate that these hosted sites or places have not yet received any reviews from customers. So imputed the missing values of reviews_per_month with a 0.

```
# we understand that reviews_per_month and last_review are null for number_of_reviews are 0.So impute reviews_per_month with 0
airbnb.reviews_per_month.fillna(0,inplace=True)
```

```
airbnb.reviews_per_month.isnull().sum()
```

```
0
```

**Checked the Duplicate rows in our dataset and no duplicate data was found**

**Identified and review outliers**

```
# outlier treatment for price:

Q1 = airbnb.price.quantile(0.10)

Q3 = airbnb.price.quantile(0.90)

IQR = Q3-Q1

airbnb = airbnb[(airbnb.price >= Q1-1.5*IQR) & (airbnb.price <= Q3+ 1.5*IQR)]
```

Similarly outlier treatment done for other numerical columns like minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count

**Data Analysis and Visualizations using Tableau**

**PPT 1**

We have used tableau to visualize the data for the assignment. Below are the detailed steps used for each
visualization.

- Preference on Room type:We identified customer preference on room type provided by listings using a bar
- Preferred Room type with respect to Neighbourhood group: We created a stacked bar chart for understanding the percentage of room type preferred w r to neighbourhood group
- Neighbourhood Vs Price and Minimum nights:we created two bar charts with avg of price and minimum nights plotted against neighbour hood to find most expensive and preferred neighbourhood

- Popular Neighborhoods: We took neighbourhood in columns and sum of reviewsin rows and took neighbourhood groups in colour.
- Location Analysis:Location wise analysis done using map to find location wise information
- Correlation Analysis:Created a correlation matrix using heatmap to find relation amoung numerical variables
- Price Vs review:plot a bubble chart where size determines reviews and color determines price

**PPT 2**

- Preference on Room type:We identified customer preference on room type provided by listings using a bar
- Customer Booking w r t minimum nights:we use side by side bars to display the distribution of minimum nights( binned ) based onthe number of ids booked for each neighbourhood group.
- Neighbourhood vs Availability Vs price:we plot horizontal bar with neighbourhood and availability as axis and include price in size and neighbourhood in color to determine highly available neighbourhood
- Top 10 price range-Top 10 price range found out based on number of reviews using histogram by binning price
- We showed top host and top locations preferred by customers using map