**The Dataset**

The dataset to be used is a CSV file named airbnb.csv, which contains data on airbnb listings in the state of New York. It contains the following columns:

* listing\_id: The unique identifier for a listing
* description: The description used on the listing
* host\_id: Unique identifier for a host
* host\_name: Name of host
* neighbourhood\_full: Name of boroughs and neighbourhoods
* coordinates: Coordinates of listing *(latitude, longitude)*
* Listing added: Date of added listing
* room\_type: Type of room
* rating: Rating from 0 to 5.
* price: Price per night for listing
* number\_of\_reviews: Amount of reviews received
* last\_review: Date of last review
* reviews\_per\_month: Number of reviews per month
* availability\_365: Number of days available per year
* Number of stays: Total number of stays thus far

Some important and common methods needed to get a better understanding of DataFrames and diagnose potential data problems are the following:

* .head() prints the header of a DataFrame
* .dtypes prints datatypes of all columns in a DataFrame
* .info() provides a bird's eye view of column data types and missing values in a DataFrame
* .describe() returns a distribution of numeric columns in your DataFrame
* .isna().sum() allows us to break down the number of missing values per column in our DataFrame
* .unique() finds the number of unique values in a DataFrame column

* sns.displot() plots the distribution of one column in your DataFrame.

**Our to do list:**

*Data type problems:*

* **Task 1**: Split coordinates into 2 columns and convert them to float
* **Task 2**: Remove $ from price and convert it to float
* **Task 3**: Convert listing\_added and last\_review to datetime

*Text/categorical data problems:*

* **Task 4**: We need to collapse room\_type into correct categories
* **Task 5**: Divide neighbourhood\_full into 2 columns and making sure they are clean

*Data range problems:*

* **Task 6**: Make sure we set the correct maximum for rating column out of range values

*Dealing with missing data:*

* **Task 7**: Understand the type of missingness, and deal with the missing data in most of the remaining columns.

*Is that all though?*

* We need to investigate if we duplicates in our data
* We need to make sure that data makes sense by applying some sanity checks on our DataFrame

**Our to do list:**

*Data type problems:*

* **Task 1**: Split coordinates into 2 columns and convert them to float
* **Task 2**: Remove $ from price and convert it to float
* **Task 3**: Convert listing\_added and last\_review to datetime

*Text/categorical data problems:*

* **Task 4**: We need to collapse room\_type into correct categories
* **Task 5**: Divide neighbourhood\_full into 2 columns and making sure they are clean

*Data range problems:*

* **Task 6**: Make sure we set the correct maximum for rating column out of range values

*Dealing with missing data:*

* **Task 7**: Understand the type of missingness, and deal with the missing data in most of the remaining columns.

*Is that all though?*

* We need to investigate if we duplicates in our data
* We need to make sure that data makes sense by applying some sanity checks on our DataFrame

To perform this task, we will use the following functions:

* pd.to\_datetime(format = "")
  + format takes in the desired date format "%Y-%m-%d"

To perform this task, we will be using the following methods:

* .str.lower() to lowercase all rows in a string column
* .str.strip() to remove all white spaces of each row in a string column
* .replace() to replace values in a column with another

The missingno (imported as msno) package is great for visualizing missing data - we will be using:

* msno.matrix() visualizes a missingness matrix
* msno.bar() visualizes a missngness barplot
* plt.show() to show the plot

To diagnose, and deal with duplicate data, we will be using the following methods and functions:

- `.duplicated(subset = , keep = )`

  - `subset` lets us pick one or more columns with duplicate values.

  - `keep` returns lets us return all instances of duplicate values.

- `.drop\_duplicates(subset = , keep = )`

### Take home question

Try to answer the following questions about the dataset:

- What is the average price of listings by borough? Visualize your results with a bar plot!

- What is the average availability in days of listings by borough? Visualize your results with a bar plot!

- What is the median price per room type in each borough? Visualize your results with a bar plot!

- Visualize the number of listings over time.

**\*\*Functions that should/could be used:\*\***

- `.groupby()` and `.agg(})`

- `sns.barplot(x = , y = , hue = , data = )`

- `sns.lineplot(x = , y = , data = )`

- `.dt.strftime()` for extracting specific dates from a `datetime` column