# Prerequisite

Download pandas library by running the following code.

In [ ]:
```
!pip install pandas
```

# Exploratory data analysis (EDA)

You will learn how to systematically approach investigating an unknown dataset while maintaining a creative and open mind to search for insights.

## Context

Airbnb is an online marketplace for people to rent places to stay.

Airbnb has rolled out a new service to help listers set prices. Airbnb makes a percentage commission off of the listings, so they are incentivized to help listers price optimally; that is, at the maximum possible point where they will still close a deal. You are an Airbnb consultant helping with this new pricing service.

## Goal

We are going to focus on a question: which features are helpful for finding out the appropriate listing price?

## Load Data

In [1]:
```
import pandas as pd
```

In [5]:
```
listings = pd.read_csv('data/airbnb_nyc.csv', delimiter=',')
```

Please check out data dictionary here

## Activities

**Q:** Can you view/print the data?

In [ ]:
```
print(listings)
```

**Q:** How large is this data?

In [7]:
```
print(listings.shape)
```

```
(30179, 81)
```

**Q:** Tell me the types of each variables using `info()` method?

In [ ]:
```
print(listings.info())
```

**Q:** Can you list all column names?

In [ ]:
```
print(listings.columns)
```

**Q:** Can you print the columns named `number_of_reviews`, `number_of_reviews_ltm`, `review_scores_rating` and `review_scores_accuracy` from rows 10000:10020?

In [ ]:
```
columns = ["number_of_reviews", "number_of_reviews_ltm",
           "review_scores_rating", "review_scores_accuracy"]

print(listings[columns].iloc[10000:10020])
```

**Q:** Can you filter/output the listings which has more than 10 reviews (`number_of_reviews` >10) AND has lower than 50 ratings (`review_scores_rating` <50)?

In [ ]:
```
my_listings = listings[(listings["number_of_reviews"] > 10) & (listings["review_scores_rating"] < 50)]
print(my_listings)
```

**Q:** Can you set index to the `listings['id']`?

In [ ]:
```
listings.set_index('id', inplace=True)
```

**Q:** Can you tabulate `self_check_in` function?

Hint: use `value_counts()` method.

In [83]:
```
listings["self_check_in"].value_counts()
```

Out[83]:
```
-1    24878
 1     5301
Name: self_check_in, dtype: int64
```

**Q:** Which features have NaN variables?

Hint: You can use `any()` function :)

In [32]:
```
features_with_nan = listings.columns[listings.isna().any()]
print(features_with_nan)
```

```
Index(['name', 'summary', 'description', 'neighborhood_overview', 'transit',
       'house_rules', 'host_since', 'host_response_time', 'host_response_rate',
       'host_is_superhost', 'host_listings_count', 'host_identity_verified',
       'neighbourhood', 'review_scores_rating', 'review_scores_accuracy',
       'review_scores_cleanliness', 'review_scores_checkin',
       'review_scores_communication', 'review_scores_location',
       'review_scores_value', 'reviews_per_month'],
      dtype='object')
```

**Q:** List your insights/takeaways from exploring this data.

The data is based on 30,179 AirBnB accommodation listings in NYC. Most of the accomodations offered in NYC are found in Brooklyn (7,024) followed by Manhattan (6,022) and then Queens (1,695). Self check-in option is offered by the small portion of the accomodations (5,301) whereas in most listings (24,878) it is not available. The accomodation price varies between $10 and $500 daily, although there is a number of listings (6) where price is set up to 0. Many listings fail to provide the summary (1,218) and/or description (604).

In [ ]:
```
nan_counts = listings.isna().sum()
print(nan_counts)
```

Please, feel free to explore more! Bonus points for adding more questinos :)

Additional questions:

**AQ1:** Print first 10 rows of the first 3 columns of the dataframe

In [ ]:
```
print(listings.iloc[0:10,0:3])
```

**AQ2:** Find the most and the least expensive accomodation in NYC

In [81]:
```
price_listings = listings[(listings["price"] > 0)]
print("cheapest accomodation:", price_listings["price"].min(),
      "| most expensive accomodation:", price_listings["price"].max())
```

```
cheapest accomodation: 10 | most expensive accomodation: 500
```

**AQ3:** Find first 10 cheapest accomodations in Manhattan using columns "neighbourhood", "description", "price" (set "price" to be > 0).

In [ ]:
```
columns = ["neighbourhood", "description", "price"]
print(listings[(listings["price"] > 0) & (listings["neighbourhood"] == "Manhattan")][columns].sort_values("price").head(10))
```

**AQ4:** Which neighbourhood has the most listings?

In [87]:
```
listings["neighbourhood"].value_counts()
```

Out[87]:
```
Brooklyn                7024
Manhattan               6022
Queens                  1695
Williamsburg            1238
Bedford-Stuyvesant      1134
                        ...
Tottenville                1
Spuyten Duyvil             1
South Street Seaport       1
Castleton Corners          1
Meiers Corners             1
Name: neighbourhood, Length: 186, dtype: int64
```

**AQ5:** How many missing values are there in each column in listings

In [107…
```
nan_counts = listings.isna().sum()
print(nan_counts)
```

```
id                          0
name                       13
summary                  1218
description               604
experiences_offered         0
                         ...
secure                      0
self_check_in               0
smoking_allowed             0
accessible                  0
event_suitable              0
Length: 81, dtype: int64
```

## References

"New York", Inside Airbnb, http://insideairbnb.com/get-the-data.html

In [ ]: