New York City Airbnb Analysis

Data Source

Source: The data was sourced from Inside Airbnb. The website is not associated with or endorsed by Airbnb or any of Airbnb's competitors; it is a collaborative/partnership website with a mission driven project that provides data and advocacy about Airbnb's impact on residential communities.

Collection: The data was compiled from the Airbnb website. Data is verified, cleansed, analyzed and aggregated. Accuracy of the information compiled from the Airbnb site is not the responsibility of Inside Airbnb; thus, due care has been taken with any processing and analysis.

Contents: The data contains public information extracted from the Airbnb website including the availability calendar for 365 days in the future, and the reviews for each listing. Full list of columns is provided below. A full list of columns is available in the Data Profile below. There are 18 columns and 43,566 rows in this data set. Accuracy of the information compiled from the Airbnb site is not the responsibility of Inside Airbnb; thus, due care has been taken with any processing and analysis.

Limitations: Some data limitations can be considered. **First**, location information for listings is anonymized by Airbnb; this means the location for a listing on the map, or in the data will be from 0-450 feet (150 meters) of the actual address. Listings in the same building are anonymized by Airbnb individually, and therefore may appear "scattered" in the area surrounding the actual address. **Second**, the Airbnb calendar for a listing does not differentiate between a booked night vs an unavailable night, therefore these bookings have been counted as "unavailable". This serves to understate the availability metric because popular listings will be "booked" rather than being "blacked out" by a host. **Third**, some hosts might not keep their calendar updated, or have it highly available even though they live in the entire home/apartment. **Fourth**, since accuracy of the information compiled from the Airbnb site is not the responsibility of Inside Airbnb, this leaves room to human error.

Ethics: No "private" information is being used. Names, listings and review details are all publicly displayed on the Airbnb site. Inside Airbnb claims "fair use" of any information compiled in producing a non-commercial derivation to allow public analysis, discussion and community benefit. All copyright and registered trademarks remain the property of their owners. With all of this into account, there are no ethical concerns/issues.

Relevancy: this data set meets the project requirements since it is open source, includes 2-3 continuous and categorical variables, has a geospatial component, and contains recent data. I appreciate the work of Inside Airbnb regarding the collection and processing of the data from Airbnb, as well as the transparency of how they operate and how what is the condition of the data. Hopefully, this project will add more insights on how this data can have a positive impact on communities.

Data Profile

Full data dictionary can be found here. This was made by Inside Airbnb.

Variable	Time Variable	Structure	Qualitative/Quantitative	Data Type
id	Invariant	Structured	Qualitative	Ordinal
name	Invariant	Unstructured	Qualitative	Nominal
host_id	Invariant	Structured	Qualitative	Ordinal
host_name	Invariant	Unstructured	Qualitative	Nominal
neighbourhood_group	Invariant	Structured	Qualitative	Nominal
neighbourhood	Invariant	Structured	Qualitative	Nominal
latitude	Invariant	Structured	Quantitative	Continuous
longitude	Invariant	Structured	Quantitative	Continuous
room_type	Invariant	Unstructured	Qualitative	Nominal
price	Variant	Structured	Quantitative	Continuous
minimum_nights	Variant	Structured	Quantitative	Discrete
number_of_reviews	Variant	Structured	Quantitative	Discrete
last_review	Variant	Structured	Qualitative	Ordinal
reviews_per_month	Variant	Structured	Quantitative	Discrete
calculated_host_listings_count	Variant	Structured	Quantitative	Discrete
availability_365	Variant	Structured	Quantitative	Discrete
number_of_reviews_ltm	Variant	Structured	Quantitative	Discrete
license			Blank	

Data Cleaning, Wrangling and Consistency Checks

Dropped columns:

'license' column dropped since it's in blank and is irrelevant to our analysis.

Deleted values: 20 entries deleted because host identity was in blank, or not clear.

Missing values: last_review and reviews_per_month have the same number of missing values. They are related to the fact that these hosts have zero reviews (both in general and also in the past 12 months). A lot of them also have no availability. Will leave it like that for now.

Renamed columns: 'id' column changed to 'listing_id' to differentiate from 'host_id'.

Replaced values: blank cells of 'last_review' and 'reviews_per_month' were replaced with 0 values in order to make the 'changing dtypes' process easier.

Converting data types: 'listing_id' and 'reviews_per_month' columns from float64 to int64 since they don't have decimals.

Key Questions:

1) How can we serve our community with Airbnb services?

2) What are the most popular neighborhoods and what is usually the price range?

3) Are price and location an important factor that customers take into consideration? If so, what is the correlation between the two and how can we improve to serve our customers better?

4) How do reviews (in general and per month) affect the host popularity?

5) What is the correlation between the minimum number of nights and availability?

6) Does the room type affect availability and demand?