Task 6.1 – Sourcing Open Data
Kaylee Michaleski

Analyzing Vancouver Airbnb Listings

Investigate the impact of listing variables such as neighborhood, room type, and minimum night stays on availability and reviews.

Data Source

Source: The data used for this assignment was sourced from Inside Airbnb. The website is not affiliated or sponsored by Airbnb – it is a website that generates data using public listing information from Airbnb's official website.

Collection: The data collected is pulled directly from the Airbnb website using only public information available. Inside Airbnb verifies, cleans, and aggregates this data for distribution on their website for the purpose of public analysis and discussion. The accuracy of the data lies with the host creating the listing on Airbnb, and the listing information in this dataset is a snapshot in time (as on June 10, 2023), rather than all available listings in the past and present. Neighbourhood names are also provided by Inside Airbnb rather than those provided on the listing, as the Airbnb neighbourhood names tend to be inaccurate.

Contents: The data contains public information pulled from Airbnb listings relating to the details of the listing such as the neighbourhood and room type, as well as calculated information relating to the reviews and availability of the listing over 365 days of the year. A full list of columns is available in the Data Profile below. There are 18 columns and 6,355 rows in this data set.

Limitations: There are a few limitations to consider when utilizing this data. As mentioned, the accuracy of the information lies with the listing creator. While it is likely a host wants their listing information to be as accurate as possible, there is room for human error. Secondly, location information for listings is anonymized by Airbnb, meaning that the true listing location will be within 0-450 feet of the posted location. The neighbourhood names in this database are compiled using the public listing coordinates, and therefore if a listing falls close to the edge of a neighbourhood, the anonymization may result in a listing categorized in a neighbourhood it does not necessarily fall into. Finally, there are limitations in the availability metric of a listing, as the Airbnb calendar does not differentiate between a listing that has been booked versus a listing that is unavailable, as in the host has blocked it off either for their own use or other purposes. It is important to keep this in mind when considering the popularity of a particular listing.

Ethics: The information in this dataset is all public information that is aggregated independently of Airbnb or any of Airbnb's competitors. Its intent has been communicated for non-commercial analysis and benefit of the community, and the collection process, policies, and assumptions have been made quite clear, so there are no immediate ethical concerns.

Relevancy & Reasoning: I believe this data set meets the necessary requirements for this project as it is open source, includes a geospatial component, meets the size and variable requirements and is recent over the past 3 years. I appreciate the thoroughness of Airbnb's collection and communication of the information and believe it sets a good example of data research. I think it will be interesting to utilize a dataset relatively close to home and gain insight on what review metrics may mean for an organization.

Data Profile

A full data dictionary can be found here, compiled by Inside Airbnb.

Variable	Description	Time Variable*	Structure	Qualitative/Quantitative	Data Type
id	unique identifier of listing	Invariant	Structured	Qualitative	Ordinal
name	name of the listing posted on Airbnb	Invariant	Unstructured	Qualitative	Nominal
host_id	unique identifier of host	Invariant	Structured	Qualitative	Ordinal
host_name	name of listing host	Invariant	Unstructured	Qualitative	Nominal
neighbourhood_group	neighbourhood group the listing is located in (blank)	Blank			
neighbourhood	neighbourhood listing is located in	Invariant	Structured	Qualitative	Nominal
lattitude	lattitude of listing	Invariant	Structured	Qualitative	Ordinal
longitude	longitude of listing	Invariant	Structured	Qualitative	Ordinal
room_type	whether the room is an entire home, private room, hotel room or shared room	Invariant	Structured	Qualitative	Nominal
price	price per night of listing	Variant	Structured	Quantitative	Continuous
minimum_nights	minimum number of nights required to book listing	Variant	Structured	Quantitative	Discrete
number_of_reviews	number of reviews received for listing in total	Variant	Structured	Quantitative	Discrete
last_review	last date of review	Variant	Structured	Qualitative	Ordinal
reviews_per_month	reviews received per month	Variant	Structured	Quantitative	Discrete
calculated_host_listings_count	number of listings a host has registered with airbnb	Variant	Structured	Quantitative	Discrete
availability_365*	the number of days the listing is available	Variant	Structured	Quantitative	Discrete
number_of_reviews_ltm	the number of reviews in the last 12 months	Variant	Structured	Quantitative	Discrete
license	the license/registration number of the listing	Invariant	Structured	Qualitative	Ordinal

The following updates were made to the dataset through data wrangling and consistency checks:

Renamed columns:

- '**id**' was changed to 'listing_id' for clarification
- '**name**' was changed to 'listing_name' for clarification

Dropped columns:

- '**neighbourhood_group'** – dropped as no data was available in this dataset; the column was blanked
- '**license**' – dropped as it is likely irrelevant to this analysis

Changed data types:

- '**listing_id'** – changed from a float to an integer data type, as an id number should not include decimals
- '**last_review**' – mixed data type was found, so the data has been updated to a string

Missing values:

- '**reviews_per_month**' and **'last_review'** columns both have missing values, but they are related to the fact that the associated listing has no reviews at all; I have chosen to leave this as is to act as an indicator of this fact

Key Questions

1. Is there a relationship between room type and availability?
2. Is there a relationship between the number of reviews in the last 12 months and the availability? Can this be used to determine which units are unavailable due to books versus host black outs?
3. Which neighbourhoods are most popular?

4. Which neighbourhoods have the highest priced listings? Is there a relationship between availability and these neighbourhoods?
5. Does the number of listings a host has affect their reviews?
6. Does the minimum number of nights required to book affect availability?