



## Toronto Airbnb Analysis

**Objective:** To build an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python.

**Data:** Summary information and metrics for Airbnb listings in Toronto, compiled June 5, 2022.

Source: "listings and calendar Toronto, Ontario, Canada", Accessed from <http://insideairbnb.com/get-the-data> on July 24, 2022.

Inside Airbnb is a mission-driven activist project with the objective to provide data that quantifies the impact of short-term rentals on housing and residential communities, as well as create a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals.

While the mission of this data source may be to highlight the negative of the Airbnb platform, the data collected has come directly from open-source data from the Airbnb website. This source has already taken sets to clean and aggregate the data, which will save time in the analysis process. It will be important to be mindful of potential negative bias in this data. The data can be verified with the direct Airbnb data if necessary.


### Data Shape and Dictionary:

Listings: 15,172 listings (rows), 18 variables (columns)

Variable Name	Description	Data Type
<b>id</b> 🔑	Unique listing ID	object
<b>name</b>	Listing name, as posted online	text
<b>host_id</b> 🔑	Unique host ID	integer
<b>host_name</b>	Host's name	text
<b>neighbourhood_group</b>	Geocoded using the latitude and longitude against neighbourhoods as defined by open or public digital shapefiles	text
<b>neighbourhood</b>	Toronto neighbourhood location	text
<b>latitude</b>	Latitude of listing location	numeric
<b>longitude</b>	Longitude of listing location	numeric
<b>room_type</b>	Listing's property type; entire home/apt, private room, shared room or hotel room	text
<b>price</b>	Daily price in local currency	currency
<b>minimum_night</b>	Minimum number of night stay for listing	integer
<b>number_of_reviews</b>	Quantity of reviews for the listing	integer
<b>last_review</b>	Date of the newest review written for the listing	date
<b>reviews_per_month</b>	Average number of reviews written per month	integer
<b>calculated_host_listings_count</b>	Number of listing the host has in the Toronto	integer
<b>availability_365</b>	Availability of the listing 365 in the future as determined by the calendar.	integer
<b>number_of_reviews_ltm</b>	Number of reviews the listing has in the last 12 months	integer
<b>license</b>	License/permit number for short term rentals	text



Calendar: 5,537,417 rows, 7 variables (columns)

Variable Name	Description	Data Type
listing_id 	Unique listing ID	object
date	The date in the listing's calendar	date
available	Whether the date is available for a booking	boolean
price	The price listed for the day	currency
adjusted_price	Adjusted day price	currency
minimum_nights	Minimum nights for a booking made on this day	integer
maximum_nights	Maximum nights for a booking made on this day	integer

### Data Assumptions:

- The source site is not associated with or endorsed by Airbnb or any of Airbnb's competitors.
- The data utilizes public information compiled from the Airbnb web-site including the availability calendar for 365 days in the future, and the reviews for each listing. Data is verified, cleansed, analyzed and aggregated.
- No "private" information is being used. Names, photographs, listings and review details are all publicly displayed on the Airbnb site.
- This site claims "fair use" of any information compiled in producing a non-commercial derivation to allow public analysis, discussion and community benefit.
- Accuracy of the information compiled from the Airbnb site is not the responsibility of Inside Airbnb. Due care has been taken with any processing and analysis.
- Location information for listings are anonymized by Airbnb.
  - In practice, this means the location for a listing on the map, or in the data will be from 0-450 feet (150 metres) 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.
- Listings can be deleted in the Airbnb platform. The data presented here is a snapshot of listings available at a particular time.
- 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.
- Some hosts might not keep their calendar updated, or have it highly available even though they live in the entire home/apartment.
  - To ensure you are only seeing listings that are both "highly available" and being booked frequently, use the "Only highly available" filter with the "Only recent and frequently booked" filter to only select listings that have only been rented recently (reviewed in the last 6 months), and are being rented regularly (number of nights per year greater than the threshold for that city).



- Neighbourhood names for each listing are compiled by comparing the listing's geographic coordinates with a city's definition of neighbourhoods. Airbnb neighbourhood names are not used because of their inaccuracies.
- All copyright and registered trademarks remain the property of their owners.

#### Data Cleansing Steps for Calendar:

1. Change id and host\_id from integers to objects as these are unique identifiers.
2. No duplicated found.
3. Missing Values:
  - a. (2) price: change to 0
  - b. (2) adjusted\_price: change to 0
  - c. (2) minimum\_nights: change to 0
  - d. (2) maximum\_nights: change to 0
4. Descriptive Analysis:
  - a. Price max of \$13,000 could be an entry error. Below is a histogram of the price distribution. A search of the Airbnb price showed a max price just under \$7,000 per night. I will mark any entries over \$7,000 as NaN for this analysis.
  - b. Minimum Nights maximum value is 2,147,483,647 days, which is an outlier. Since I cannot verify this data, I will limit the analysis to 90 days by removing the values over 90 days
  - c. Minimum Nights maximum value is 1,125 days, which is about 3 years. Since I cannot verify this data, I will limit the analysis to 90 days by removing the values over 90 days

#### Data Cleansing Steps for Listing:

1. Change price, adjusted\_price, minimum\_nights, and maximum\_nights from objects and floats to integers.
2. No duplicated found.
3. Missing Values
  - a. (2) name: ignore
  - b. (3) host\_name: ignore
  - c. (15171) neighbourhood group: no data in this column, remove the column.
  - d. (3223) last\_review: assumed to be listings with no reviews.
  - e. (3223) reviews\_per\_month: assumed to be listings with no reviews.
  - f. (9475) license: assumed to be unlicensed listings.
4. Descriptive Analysis:
  - a. Price max of \$13,000 could be an entry error. Below is a histogram of the price distribution. A search of the Airbnb price showed a max price just under \$7,000 per night. I will mark any entries over \$7,000 as NaN for this analysis.
  - b. Minimum Nights maximum value is 1,125 days, which is about 3 years. Since I cannot verify this data, I will limit the analysis to 90 days by removing the values over 90 days
  - c. Minimum Nights maximum value is 1,125 days, which is about 3 years. Since I cannot verify this data, I will limit the analysis to 90 days by removing the values over 90 days



**Questions to Explore:**

1. Does the number of reviews affect the price?
2. Do hosts with multiple listings have more reviews per listings?
3. How does accommodation type affect the price?
4. Are listings with less availability priced higher or lower?
5. Do hosts with multiple listings have less availability per listing?