# Towards an Automated Categorization and Rating System of Airbnb Listings in New York City

Jonathan Pichot, Fernando Melchor, and Avikal Somvanshi
Center for Urban Science + Progress
New York University
New York, NY

***Abstract*—**

## I. Introduction

Airbnb is an online platform for residents of cities around the world to rent space in their homes and apartments. Founded in 2008, Airbnb's mission is to help people "monetize their extra space." They've been very successful, with over 3 million listings in over 65,000 cities worldwide.[1] Since Airbnb listings are short-term housing provided by residents, the quality of the space can vary drastically in quality, appointment, and amentities. The primary way Airbnb differentiates its housing options is by the nature of the room. There are three options: shared, private, or entire home. A shared listing is a shared space with the host, often on a couch in a living room. A private listing has a door, usually a bedroom. And finally, a host can rent their entire home or apartment for a period of time.

Airbnb allows hosts to list the attributes of their listing, including photos, access to technology, parking, washer/dryer, and other amentities. Past guests can leave reviews that give further context to the quality of the place. But Airbnb does not rate the neighborhood or location of the listing itself. For some more popular destinations, Airbnb does provide neighborhood guides on their website.

This project will explore the possibility of creating a machine-learning driven categorization system of Airbnb listings in New York City. A rating system similar to the 'star rating' systems developed by the departments of tourism in several countries is a good example of how this kind of rating system might work. This project will focus on neighborhood attributes around a listing rather than attributes on the listing itself, but the techniques used could be applied to listing attributes as well.

The classification system developed could be useful to tourists and Airbnb alike. Using the system, Airbnb customers in New York City would get a better idea of the amenities available in the neighborhood around their rental, and Airbnb could use the analysis to better understand their listing inventory and customer preferences. Is there stronger demand for cheaper listings? For listings with better public transit connectivity? For listings close to certain kinds of amentities? There are several rich possible applications.

## II. Data and Methods

### A. Airbnb Listings

The Airbnb listings for New York City were collected from InsideAirbnb.com,[2] a website run by a New York City-based housing activist named Murray Cox. The website scrapes Airbnb's website for many cities around the world, creating snapshots of all listings on the site in a city on a given day. The data used in this analysis was from the scrape of New York City on March 2nd, 2017.

The scraped data includes a lot of relevant information that was used in the analysis including the price of the listing, how many reviews have been posted on it, its approximate location (Airbnb does not publish the exact location of listings for security reasons), the minimum number of nights per booking, the room type (shared, private, entire), and many more.

### B. Outliers

To make sure the analysis was performed on Airbnb listings that are actually being rented we removed certain outliers. This left us with the listings with the below attributes:

- Minimum of 7 or less nights per booking
- Listing price of $500 or less
- At least 1 review

We feel a listing requiring a booking of greater than 7 nights begins to be considered a sublet rather than short-term housing similar to a hotel room. Some listing had absurdly high listing prices. $500 a night is the equivalent of a high end hotel. Finally, by requiring at least 1 review we removed listings that had likely never been booked. After removing all outliers we end up with 28,970 listings.

### C. Custom Attributes

In addition to the attributes collected by Inside Airbnb, we added four custom attributes to each listing. We developed using publicly available data. They are:

- Median Household Income
- Craft Beer Count
- Specialty Coffee Count
- Connectivity Score

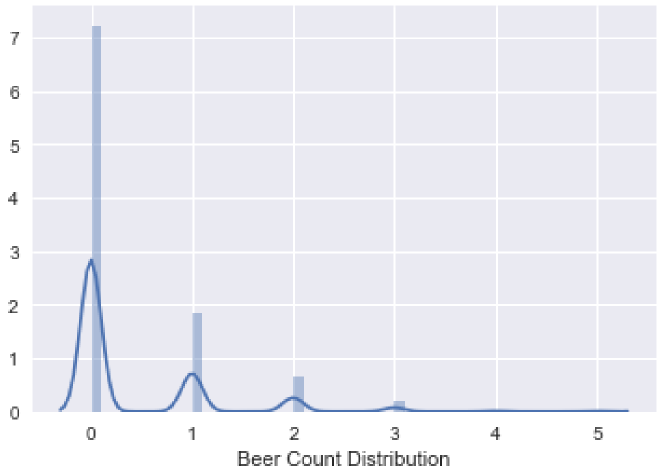

Fig. 1. Distribution of Beer Count

### D. Median income

### E. Craft Beer and Specialty Coffee Counts

One way to disguinsh a neighborhood is to identify the kinds of businesses it can support. Certain kinds of businesses cater to certain tastes. Two kinds of establishments that have become popular in what could be called the 'tastemaking young professionals' class is speciality coffee (also referred to as third-wave coffee) and craft beer. The density of these kinds of establishments in a neighborhood should work as a good indicator of the kind of clientele a certain neighborhood attracts.

The location of speciality coffee shops and breweries were collected using Yelp's API.[3] All coffeeshops in New York City that were returned from the API when searching for the string 'third wave coffee' were collected. Similarly, all breweries that were returned from the API when searching for 'brewery' were also collected. This resulted in lists of 375 coffeeshops and 67 breweries.

These lists were then run through Mapzen's Isochrone API.[4] This endpoint takes a point and returns a polygon that represents all the area one can travel to given a certain time and using a certain mode. Every coffeeshop and brewery was then merged on a polygon that represented the area one could access in 10 minutes while walking (also known as a walkshed).

Finally, using these walksheds, the total number of speciality coffeeshops and breweries within a 10 minute walk was calculated for every Airbnb listing. These totals gave us what we called our 'coffee count' and 'beer count'. The distributions of these attributes can be seen in Figure 3 and Figure 1 respectively.

### F. Connectivity Score

## III. Clustering

We

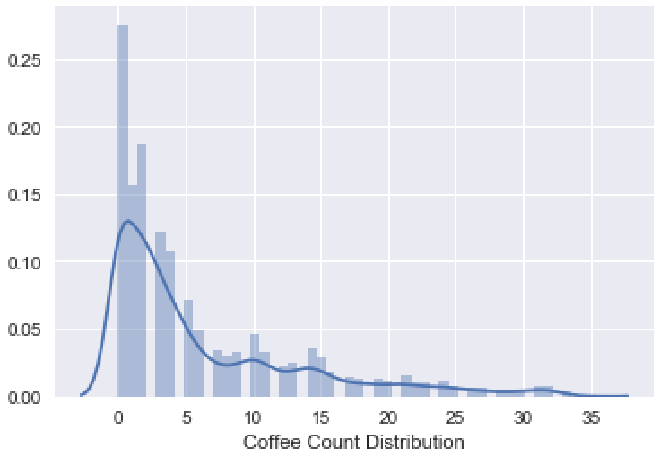

Fig. 2. Distribution of Coffee Count

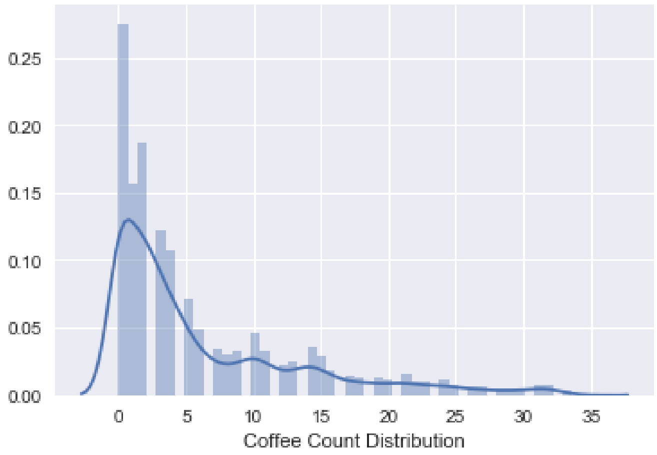

Fig. 3. Asthma Hospitilization Rate per 10,000 residents

### A. Neighborhood Profiles

## IV. Discussion and Conclusion

## References

[1] "About Us - Airbnb." [Online]. Available: https://www.airbnb.com/about/about-us

[2] "Inside airbnb. adding data to the debate." [Online]. Available: http://insideairbnb.com/about.html

[3] "Yelp fusion api." [Online]. Available: https://www.yelp.com/developers/documentation/v3

[4] "Mapzen: Isochrone api." [Online]. Available: https://mapzen.com/documentation/mobility/isochrone/api-reference/