Data 601 Project Proposal

SEPTEMBER 19, 2022

IMAD AHMAD (30198988), IBTASSAM RASHEED (30201479), YIP CHI MAN (30183603)

Introduction

Airbnb® is an American company operating an online marketplace for lodging, primarily for vacation rentals. It has grown in popularity quite rapidly; as of 2021, Airbnb® contains over 7 million listings worldwide, with a listing growth rate of over 100% (Dogru et al., 2021). Airbnb® has led to the inception of several companies supplying a similar service, such as *Vrbo*®, *Sonder*®, and *Homestay*® (Guttentag, 2013).

With the rise of the shared economy and more individuals choosing independent homestays over hotels for vacationing, shared accommodations are the way of the future, with Airbnb® leading the pack. Thus, it is important to understand the most influential metrics on Airbnb® bookings, as it gives us insight into what individuals prioritize when booking a homestay. **The purpose of our study is to perform an exploratory data analysis of a dataset containing over 250,000 Airbnb® listings across 10 major cities (Bhat, 2021)**. We aim to use data visualization to gain a deeper understanding of the impact different metrics have on our various variables of interest. Some of the questions we hope to answer include: how do seasonal Airbnb® booking trends vary over different geographic locations, and which metric is the best predictor of a listing being booked, as well as getting a strong review.

Guiding Questions

Below are the guiding questions that we are going to explore with our data analysis, both quantitatively and visually. We will attempt to gain insights into the different metrics that play a role in bookings, booking seasonal/geographic trends, and host quality.

Note: Some of the questions involve us looking into the number of times a listing has been booked, a metric that is not included in our dataset. We will use the number of reviews a listing has to quantify how many bookings it has. This should not be a problem, as we are only comparing the number of bookings for a listing relative to other listings, and not looking at the actual number of bookings.

1. **What combination of metrics (ex. Price, location, name, etc.) is the best predictor of an Airbnb® listing being booked?**

Which columns in the data were most strongly correlated with a high number of listings for a booking? Is there one specific metric that is a good predictor, or should metrics be combined to predict bookings? This will help us gain insight into what customers look at when booking.

2. **Have Airbnb® seasonal and geographic booking trends been the same every year, including the years impacted by COVID-19?**

COVID-19 initially reduced the number of vacations people were taking. However, as time went on and people began to travel again, were the same historic cyclical travelling trends seen in major cities? This will help us gain insight on travel patterns post covid, which can further be studied by comparing if the same travel patterns exist with hotel bookings.

3. **Do Airbnb® listing and booking trends vary geographically?**

Vacation trends follow a cyclical seasonal cycle. Does this same cycle exist for all major cities, or do other factors play a role in a specific city's booking cycle (ex. Climate, Location, etc.)? Do variables such as "Amenities" play the same role in predicting listing success in every major city (i.e., Is a specific amenity valued in one city over another, or are they valued equally in every city)? This will help us gain insight into whether customer listing priorities change depending on where they are travelling.

4. **What metrics predict if a host will be a "good host"?**

We will quantify being a "good host" with several metrics, such as if the host is a super host, the host response rate, etc. We will investigate which column(s) correlate most strongly with these "good host" metrics. This will help us gain insight into whether Airbnb's® metrics on what a good host match what customers consider a good host.

Dataset

The dataset was sourced from Kaggle (Bhat, 2021), and was originally pulled through Airbnb's® API (Airbnb, 2022). We have permission to use the dataset through CC0 1.0 public domain dedication (CC0 1.0, 2022) , as it is open public domain data with no copyright. We will be looking at two data tables in total:

- One contains information on 250,000+ Airbnb® listings across 10 major cities. This table has 33 columns
- The other contains 5 million+ reviews on these listings from 2008-2021. This table has 4 columns

Both data tables are structured and tabular, and contain different data types including:

- Categorical (e.g., property type, room type)
- Numerical (e.g., number of bedrooms, price, review score)
- Date (e.g., review date, host since date)
- Boolean (e.g., host is superhost, host identity verified)

The "listing ID", common to both tables, will be used to join them. In terms of the types of columns, we have:

- Listing information:
  - Listing ID, city, district, neighborhood, latitude, longitude, property type, room type, accommodates, bedrooms, amenities, names, price, min. nights, max. nights, instant bookable
- Host information:
  - Host ID, host since, host location, host response time, host response rate, host acceptance rate, host is superhost, host total listings, host has profile picture, host is verified
- Review information:
  - Review score, accuracy score, cleanliness score, check-in score, communication score, location score, value score, review ID, review date, reviewer ID

Tasks

The data cleaning phase of the two data tables will begin by changing the data types of the columns to appropriate data types. Many of the data types are "object" now, and we will have to convert them accordingly to work with them. We will assign suitable column names and identify missing values in each column. The missing values in each column will be dealt with on a case-by-case basis depending on the right way to clean and filter that specific column. Many of the columns containing IDs have varying numbers of digits, so we will add zeroes accordingly to make them all the same length. We will also utilize the second data table titled ('Reviews.csv') from the same source which will allow us to compare based on review count. Text columns like 'Listing title' or 'Listing description' will give us opportunity for frequent keywords' identification.

After data cleaning, we will move to exploratory data analysis. We will make use of Matplotlib library to visualize the clean data and find patterns and outliers. Other libraries learned in the upcoming weeks will also be utilized for this analysis including NumPy and Pandas. Studies attempting to calculate similar metrics have used these visualizations successfully (Dhillon, 2021; Sinthong, 2021):

- Bar charts and Area plots to see number of hosts over a period
- Histograms and Pie charts to see distributions of location, property type and other parameters of a listing
- Scatter plots and Line charts to check correlations between the number of bookings and other metrics; as well as to see what metrics correlate with "good host" metrics

Finally, we will verify the sequence of code and visuals so that the data story can be told in an efficient and captivating manner. We will clean the visuals and make sure everything is laid out in a logical way that can be easily understood. We also plan to brainstorm a way to make the presentation interactive so the class can feel engaged.

It is difficult to determine a breakdown of tasks at this time, as we are not fully sure what will be required for each of the tasks. We will attempt to work on each task together so we can all have a strong understanding of the full data visualization process. As a soft breakdown, Edmund will focus on the cleaning, Ibtassam will focus on the visualization, and Imad will focus on what data questions to ask, as well as what the best way to present is.

References

Airbnb, 2022, *Airbnb API*

Bhat, M. 2021, *Airbnb Listings & Reviews*, electronic dataset, Kaggle, viewed 13 Sept. 2022, <https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews>

CC0 1.0 Universal Public Domain Dedication, 2022. *Creative Commons*, <https://creativecommons.org/publicdomain/zero/1.0/>

Dhillon, J., Eluri, N., Kaur, D., Chhipa, A., Gadupudi, A., Eravi, R. and Pirouz, M., 2021. Analysis of Airbnb Prices using Machine Learning Techniques. IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp.0297-0303.

Dogru, T., Mody, M., Line, N., Hanks, L., Suess, C. and Bonn, M., 2021. The Effect of Airbnb on Hotel Performance: Comparing Single- and Multi-Unit Host Listings in the United States. *Cornell Hospitality Quarterly*, 63(3), pp.297-312.

Guttentag, D., 2013. Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, 18(12), pp.1192-1217.

Polisetty, A. and Kurian, J., 2021. The Future of Shared Economy: A Case Study on Airbnb. *FIIB Business Review*, 10(3), pp.205-214.

Sinthong, P. and Carey, M., 2021. Exploratory Data Analysis with Database-backed Dataframes: A Case Study on Airbnb Data. *2021 IEEE International Conference on Big Data (Big Data)*, pp. 3119-3129.