AirBnb Database System Project Description

Simao (Alice) Chen
4916@nyu.edu

While battling against Covid-19, the world is gradually opening up for tourism. Having memorable trips through AirBnb is getting increasingly popular among people around the world. Airbnb does not only provide affordable trips but also connects the world together through cordial hosts and unique listings. This application is9 meant to create a database system for AirBnb hosts and rentals and other relevant entities. The users of this application can be data scientists who are interested in what kind of hosts and listings are active on this website, or prospective customers who intend to book some rooms to stay for leisure or business trips. All the data is made up, but I borrow countless inspirations from the actual dataset by AirBnb.[1] The entities' attributes and the entities' relationships are closely modeling what is happening in the real world with AirBnb. I use my own personal experience booking AirBnb services in the past and policies provided on AirBnb to model the entity sets and relationships. It brought me joy and happy memories, and I hope users can have the same experience.

Entities, relationship sets, and business rules:

1. **Hosts:**
   A host can provide listings. A host may be able to provide as many listings as they want.
   A host is uniquely identified by **hid**.
   A host must reside in exactly one neighborhood.
   A host shares at most one response_info that describes them.
   A host is described by a name, and whether they have pet(s).
   Everyone can register a host profile page online and become a host. However, a host may not create their listings right when they create their profile. Once AirBnb obtains necessary info about a person, they are a host. A host may also take a break and therefore do not have listings on AirBnb, but it does not change their host status.
2. **Listings: (Strong entity to rooms)**
   A listing must have a host and have at most one host, must contain at least one room, and is located in exactly one neighborhood.
   A listing is uniquely identified by **lid**.
   A listing has at most one rating.
   A listing is described by property_type, a price (in dollars per night, per listing).
3. **Neighborhoods:**
   A neighborhood is uniquely identified by **n_name**, and is described by whether there is a restaurant, a museum, and curfew at night.
4. **Rooms: (Weak entity to Listings) (Strong entity to Availability_dates)**
   A room must be contained in exactly one listing. Several rooms can belong to one listing.

[1] "Get the Data - inside Airbnb. Adding Data to the Debate." *Get the Data*, Get the Data, 7 Sept. 2021, http://insideairbnb.com/get-the-data.html.

If a listing is taken off, the room(s) is also taken off.
A room is uniquely identified by a tuple **(lid, room_id)**.
Room_type describes a room as either a private room, shared room, or the entire place.
A room also describes Num_people that can stay in.
A room may have an availability_dates.

5. **Availability_dates: (Weak entity to Listings)**
   An **Availability_dates** is uniquely identified by **a tuple (lid, room_id, start_date, end_date)**, The start_date and the end_date indicates when the availability_dates lasts.
   Every availability_dates must correspond to exactly one room.
   Some rooms may not have availability_dates.
   If the room is taken off, the availability_dates is taken off.
6. **Rating:**
   A rating describes exactly one listing, and some listings may not have a rating.
   A rating is uniquely identified by **rate_id**.
   Every rating has an average_score (a decimal out of 10).
7. **Response_info:**
   A response_info evaluates exactly one host, and some hosts may not have responded to any prospective customers yet and therefore do not have response_info.
   A response_info is uniquely identified by **res_id**.
   Response_time is the average time it takes for a host to respond to all customers' questions or requests.
   Response_rate is the percentage a host responds to all the received messages.
   A response_info is described by response_time and response_rate.
   A response_info is meant to evaluate how responsive and engaged a host is in managing the listing. Guests may get a good idea of how soon they can confirm their booking and expect how frequently they can communicate with the hosts.

The way I model the actual data to create made-up data is to read and examine the original dataset provided by AirBnb, especially the headers and some interesting relationships. Due to privacy concerns, the dataset does not provide all information and sometimes does not represent the relationships between entities. I create data that represent the common scenarios in real life and how hosts manage listings and users interact with hosts on AirBnb.

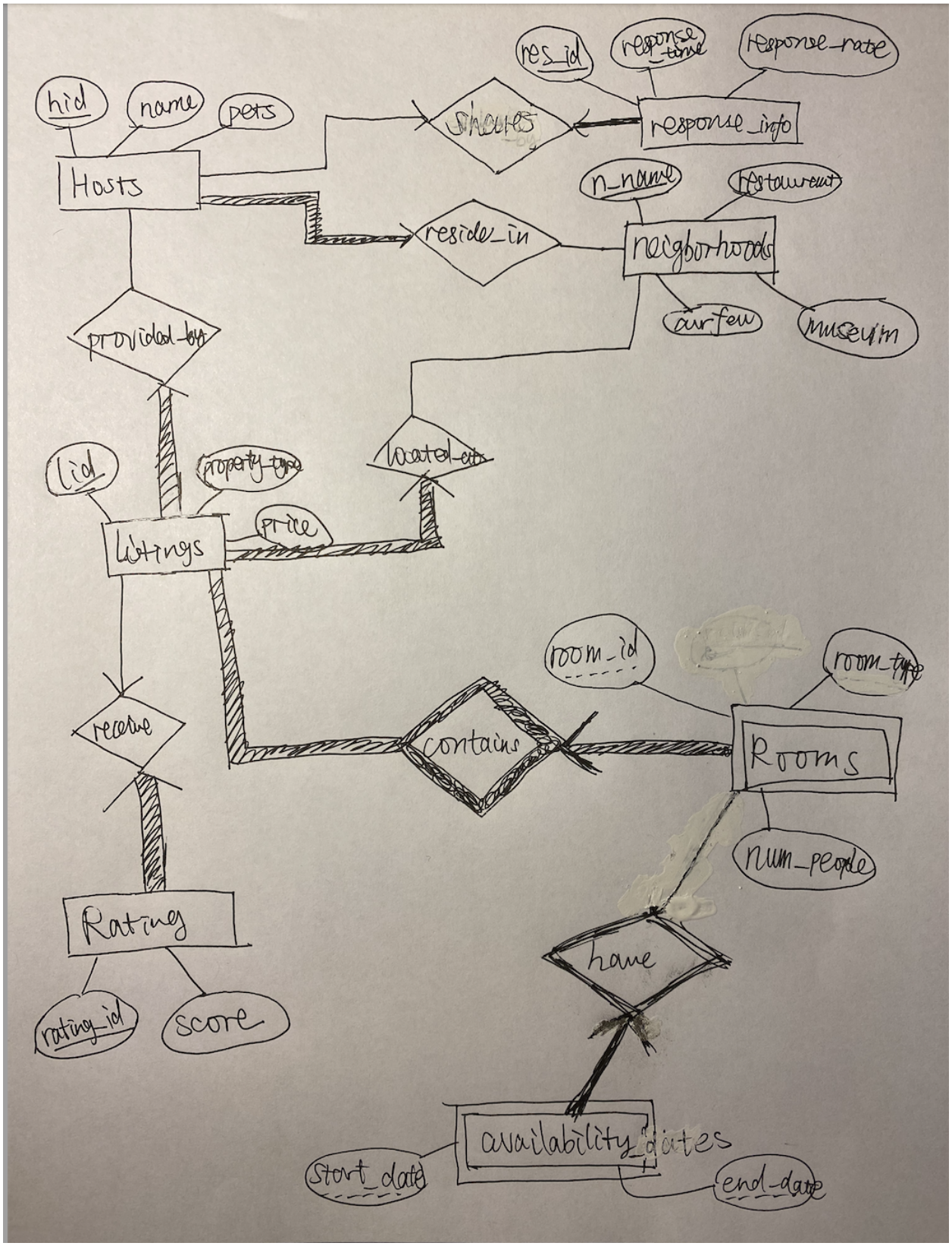
After the database is connected to a website, the users should be able to type in SQL queries that will help them understand the dataset better. Some of those questions could be:

1. Show the names of the host, the property_type of the listing, and the names of the neighborhood, where the listing has 2 rooms that can fit in 3 people in total and are available during September 01-05, 2021, and the listing must be located in a neighborhood that has a restaurant and no curfew.
2. Show the names of the host, the property_type of the listing, and the name of the neighborhood, where the listing is below $100 per night, has 5 rooms that can fit in 5 people in total. The listing should not have pets, as the guests are allergic, and should

be located in a neighborhood where there is a restaurant, a museum, and must have curfew.

3. Show the names of the host, the property_type of the listing, and the prices, where the neighborhood must have a bar, and is available between September 20-26th, 2021, and the rating must be above 8, and the hosts' response_time must be within an hour.
4. Show the number of available listings during the days between September 25th to September 29th, 2021, where the hosts must have a response_rate of 100% and response_time within 2 hours, and the rating must be above 7.
5. Show the number of hosts who reside in a different neighborhood from the listings' neighborhoods that they are providing.

I attached the ER Diagrams in the next pages.


res_id
response_time
response_rate
hid
name
pets
shows
response_info
Hosts
n_name
restaurant
reside_in
neighborhoods
curfew
museum
provided_by
located_at
lid
property_type
price
Listings
room_id
room_type
receive
contains
Rooms
num_people
Ratings
have
rating_id
score
availability_dates
start_date
end_date