2017-12-09

FINAL REPORT

TEAM MEMBERS:

YANLIN YU | JIALEI GUO | YIYI YE | RAN ZHOU | ZHAOXI ZHANG | JIAHUI WANG

1. DESCRIPTION OF THE APPLICATION

Our application is a data-based web-page design, which provides a two-way service for both house hosts and renters in Toronto according to their preferences. For hosts who want to lease their houses, we have several criteria for them to input. To be more detailed, hosts need to input the location of their house, the number of guests the house can contain, the room type of the house(entire room/private room/shared room), whether the house is instantly bookable, the number of rooms, beds and bathrooms, amenities the house provides(TV, breakfast, washer, shampoo, air conditioning etc.) and the house rules (pets allowed, smoking allowed, suitable for events etc.). Based on the information provided by a host, we select other hosts who have the similar conditions to calculate the average price of these hosts and offer it to the host as a referenced price. For the renters, they also have different criteria for choosing the house, like location, time range of renting, number of people, expected price range. We add three filters into our website: crime rate, shopping and food, which can be used by renters as filter elements. Based on these preferences, the web-page will offer several ranked available choices and the link containing detailed information for each house.

2. DATA SELECTION

We use three real datasets for our application:

1) Airbnb hosts' data in Toronto
   (http://insideairbnb.com/get-the-data.html)

2) Crime rate data in Toronto

(http://data.torontopolice.on.ca/datasets/mci-2016/data?geometry=-80.368%2C43.378%2C-76.412%2C44.073)

3) Yelp data in Toronto

(https://www.yelp.com/dataset/documentation/sql)

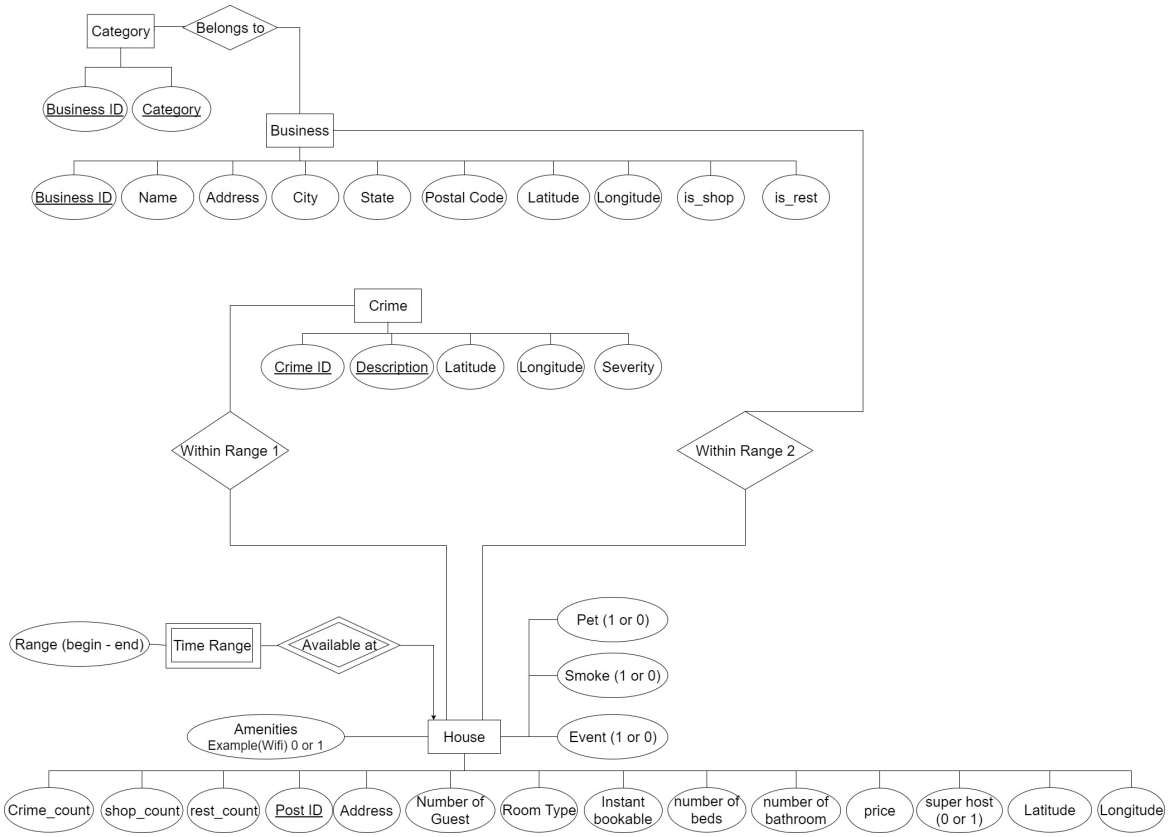
For Airbnb host dataset, we delete several descriptive variables and also the ones that have slight influence on the house price. We also delete the observations that have lots of missing values. The same work also did for the yelp dataset.

For crime rate dataset, we just pick several useful columns, i.e., unique id, type of MCI, longitude and latitude which can be connected to the house location and then further impact the renters' choice of house, plus the price the host will quote .

3. ASSUMPTIONS

- Only within a certain distance will the crime rate and yelp business affect the rental choice from the user.
- When other conditions are the same, users will have the same preference for two houses close in certain distance, i.e., the situation that the house on one street is significantly more preferable than the one on the adjacent street is not considered.
- No houses in the database can be rented or reserved until the website is accessible, i.e., all the houses will be available
- No booking activity have been created, so all the houses in the dataset
- Crime severity will be the only security concern for the user.
- The safety rating of one area is independent of when crime happens.
- Policies of smoking, pet and event are the only three house rules people are concerned about. Other house rules will not be considered as the criterion for rental decision. The similar assumptions are applied to the amenities of a house as well.

4. E/R DESIGN



5. DATABASE TABLES

- **business**(id, name, address, city, state, postal_code. latitude, longitude, stars, review_count, is_open, is_shop, is_rest)
- **category**(business_id, category)
- **businessBelongsToCategory**(business_id, category) --> which can be merged with category
- **crime**(crime_id, location, description)
- **House**(id, longitude, latitude, guest_num, room_type, price, instant_bookable, bedrooms, beds, bathrooms, superheats, pet_allowed, smoking_allowed, Free_parking, Family_kid_friendly, Washer, Hangers, Lock_on_bedroom_door, Wireless_Internet, Laptop_friendly_workspace, TV, Shampoo, Self_Check_in, Heating, Hair_dryer, Indoor_fireplace, Dryer, Iron, Breakfast, Doorman, Buzzer_wireless_intercom, Air_conditioning, Pool, Kitchen, Crime_count, shop_count, rest_count)
- **Time_Range**(house_id, start_date, end_date)
- **AvailableAt**(house_id, start_date, end_date) --> which can be merged with time_range
- **businessWithinHouse**(business_id, house_id)
- **crimeWithinHouse**(crime_id, house_id)