Yasmine Boukataya, Gabriela Campoverde, Yiyasu Paudel, Purva Sheth
CIT 550
Project Proposal

Boston AirBnBs and Weather

For our project we will create an application which evaluates the relationship between AirBnBs and the weather for the Boston area. We gathered two datasets from Kaggle which are specific to Boston. There are 3 tables in the boston Airbnb dataset: listings, calendar, and reviews. Listings contains information about a particular listing such as its name, id, summary, and a url to the listing on airbnb.com. Calendar contains the listing id and the date that it was booked.[1] There is 1 table in the Weather dataset which contains the date, temperature, dew point, and humidity.[2]

Our target user for our web application is a person who lives in Boston and is considering becoming an AirBnB host. Our application will give a user the opportunity to query through Boston listings and assist them in how this user should price their rental according to how people price their listings in their surrounding area. Moreover, the user will be able to look up the ratings and other host and listing statistics so that the user can be well informed about how competitive the area is. We will marry this AirBnB dataset with a Boston weather dataset to allow users to evaluate the effect of weather on the availability and price of listings in the area. Boston is known for its cold, early winters, and our tool will allow these users to measure the effect of weather on AirBnB rental demand.

Summary Statistics

To grasp an overview of the datasets we evaluate the Weather and AirBnB data sets separately for the following high-level metrics:

1. **Weather:**
    a. Columns : Year, Month, Day, High Temp (F), Avg Temp (F), Low Temp (F), High Dew Point (F), Avg Dew Point (F), Low Dew Point (F), High Humidity (%), Avg Humidity (%), Low Humidity (%), High Sea Level Press (in), Avg Sea Level Press (in), Low Sea Level Press (in), High Visibility (mi), Avg Visibility (mi), Low Visibility (mi), High Wind (mph), Avg Wind (mph), High Wind Gust (mph), Snowfall (in), Precip (in), and Events
    b. Statistics of weather dataset:
        i. 3749 rows × 24 columns (all columns except the last one have numerical values). There are no ‘NULL’ in the dataset

[1]Boston AirBnB Dataset: https://www.kaggle.com/airbnb/boston/metadata
[2]Boston Weather Dataset: https://www.kaggle.com/jqpeng/boston-weather-data-jan-2013-apr-2018

ii. Summary Statistics on some columns (column name, mean, standard deviation, minimum, maximum)

1. Year, 2012.640437, 2.966, 2008, 2018
2. Month, 6.41, 3.47, 1, 12
3. Day, 15.69, 8.807, 1, 31
4. High Temp(F), 59.53, 18.35, 12, 103
5. Avg Temp (F), 52.37, 17.36,2,92
6. Low Temp(F), 44.706, 16.835, -9, 81

2. Airbnb

a. Calendar table: 33.52 MB, 4 attributes, 1,309,000 rows
Summary Statistics (column name, mean, standard deviation, minimum, maximum) :
(date column, 6 March 2017, 1 month 34 days, 5 Sep 16, 4 Sep 17)
(available column boolean, no mean std dev etc)
(price column, 143.8293, 38, 65.00, 1050.00)

b. Listings Table: 14.76 MB, 95 attributes, 3585 rows
Summary statistic for some attributes we plan to use, there are over 95 attributes so impossible to include all of them
(column name, mean, standard deviation, minimum, maximum)
(host_response_rate, 76.0283%, 11.383%, 0%, 100%)
(host_acceptance_rate, 60.399%, 0%, 100%)
(host_listing_count, 78.041,30.392, 0, 749)
(cleaning_fee, 84.211, 32.383, 5, 150)

c. Review Table: Size: 27.32 MB, 6 Columns: listing_id, id, date , reviewer_id, reviewer_name, comments 68276 Rows (Including header)
Summary Statistics: No null values for listing_id, id, date and reviewer_id. However, some reviewer_name and comments are unclear. Dataset includes reviews from the year 2009 - 2016 and the count is monotonically increasing. Maximum review is from the year 2016. 63789 unique reviewer id out of 68275. Hence, 4486 reviews are from returning customers. No statistical analysis for reviewer_name and comments as the value is non numeric.

Queries

In order to service the different use cases described above, we will include queries like the following:

1. We will need an aggregate query to get the average temperatures (High, Low and Average) for a particular day considering the temperature on that day in the previous years.
2. We can provide the average listing ratings for a set time frame for Boston AirBnBs.
3. We can provide the average price of a listing for a set time frame for Boston AirBnBs.

4. We can create a date column in the weather_clean table using the values from Year, Month, Day and then join the date column in the calendar table. Example below:

```
WITH weather AS (SELECT CONVERT(date,CAST([Year] AS VARCHAR(4))+'-'+ CAST([Month] AS VARCHAR(2))+'-'+ CAST([Day] AS VARCHAR(2))), avg_temp AS temp FROM weather_clean)
SELECT D.date, W.date, W.temp
FROM date D, weather W
JOIN ON D.date = W.date
```

5. We can include the review from the reviews table by joining it on the date column in the calendar table to know more about peak/off seasons which may be helpful for future customers to plan a trip accordingly.

With the above datasets and queries, we believe that we can create a great web application to inform users who are interested in becoming AirBnB hosts in Boston.