Xi Lin

Network Science (6201) – Summer 2021

Abstract

To propose a method for using graph to analytics to identify and target the best reviewed listing of \*\* city in term of price group and neighborhood. Providing visitors with the most effective information to help select the best quality short time rental properties at the psychological price point and tourist location

**USE GRAgh ANALYTICS TO INDENTIFY THE BEST REVIEWED LISTINGS OF cAMBRIDGE, mA USA IN TERM OF PRICE GROUP AND NEIGBORHOOD.**

Project Proposal

# Primary Objective:

Airbnb, Inc. is an American company founded in 2008 that operates an online marketplace for lodging (primarily homestays for vacation rentals) and travel activities. The platform is based in San Francisco, California, and is accessible through its website and mobile app. Airbnb does not own any of the properties listed; instead, it profits by collecting a commission from each booking. The emergence of Airbnb has contributed to a boom in tourism and improved the quality of travel.

This proposal is to apply graph query to identify and target below:

1. Rental properties that belong to same host/owner;
2. Reviews that are coming from same reviewer;
3. The best reviewed listings of Cambridge city in term of price group and neighborhood. Usually the visitors or travelers are looking for the best available rental properties either by price range or location/neighborhood. This can be done via SQL query by filters and orders, however, exploring graph could end up with more intuitive views and single hops.

# Dataset & Business Use Case:

Data was acquired directly from [Airbnb’s database](http://insideairbnb.com/get-the-data.html) for educational research purposes. The dataset we use in here is the listings and reviews data that is published by Airbnb per period per region. In US, the region usually is a city. Based upon the data volume requirement, we deliberately choose a town that has less data but good enough for demo purpose – we pick the listings and reviews data from Cambridge, MA USA as of March, 2021.

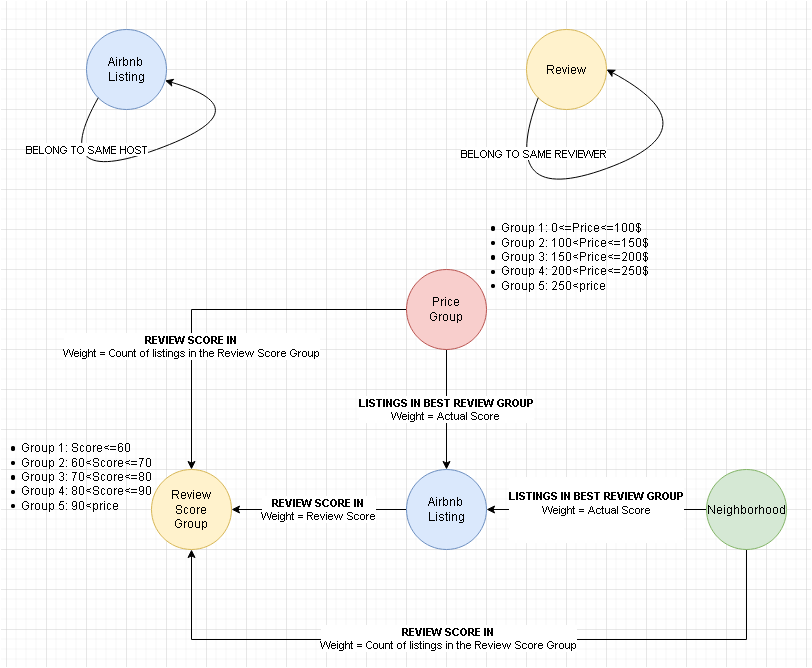
A general overview of the data set as follows (a dictionary is included in the Appendix for deeper review):

* **Listings / Listing Detail:** Listing data for Cambridge, MA in terms of link, rental type, number of bedroom and restroom, accommodation, neighborhood, review counts and average ratings, default price, so on so force.
* **Calendar:** Despite of its misleading name, this data provides the daily price (could vary) info upon each listing (a short window starting from March 2021).
* **Reviews / Review Detail**: Provides the available historically review info upon each listing, it has listing id, date, reviewer, and comments. Surprisingly no rate score – this is for privacy purpose.
* **Neighborhood**: this is a reference data set for neighborhood.

Overall, this is public data that doesn’t have a lot of financial or market data – hence we don’t see the actual rent-out records along with the past price information. Instead, we can see the listing information, and the accumulated review information.

For our project, we don’t need all of them – the core dataset is the Listings / Listing Detail and Review/Review Detail. We don’t even need all the attributes - The columns we need in Listings are listing id, host id, price, neighborhood, number of reviews, and review scores rating. In Reviews we need review id, reviewer.

# Graph Data Model



# Data preparation and cleaning：

1. Download data from Airbnb data set: <http://insideairbnb.com/get-the-data.html>
2. Filter out the listings with no price – No price could mean not available or the host may take this property off, default its price to zero makes no sense.
3. Filter out the listings with no review – we can’t default it to zero either.
4. Pre-calculate the best review rating score per each Price Group and per each neighborhood so we can make the edges between

* Price Group and Review Score Group
* Neighborhood and Review Score Group

1. Pre-calculate the listings in the best review score group per Price Group and per Neighborhood so we can make the edges between

* Price Group and the Airbnb Listing
* Neighborhood and the Airbnb Listing

# Planned Analysis Techniques：

## (1) Install the graph database Neo4j desktop and create local graph database; A quick hands-on training is as below.

* [Hands-on Introduction to Neo4j](https://message.neo4j.com/NzEwLVJSQy0zMzUAAAF9pQWIT9sGO0JB6zOg64no5-RYjzLeZjxcw0OSufgcFnaQQc5sT7OEzYYWksUS1r8BBAF4v9M=)
* [Hands-on with Neo4j Aura Free Tier](https://message.neo4j.com/NzEwLVJSQy0zMzUAAAF9pQWITw7ibIOVjnRsZkKLhdLEovWkSZ5YfccjOgkr6hm0tJXhBeSg5n1NC57mXcers9ZzDl0=)
* [Getting Started with Neo4j Bloom](https://message.neo4j.com/NzEwLVJSQy0zMzUAAAF9pQWIT4qafJT1nbYfv1i8Cdmo04Rzs3VdO86IGTbKnynAuocMdVlwlWcYMtunnATvXlcJUkE=)
* [Building GraphQL APIs with the Neo4j GraphQL Library](https://message.neo4j.com/NzEwLVJSQy0zMzUAAAF9pQWIT5B6ZWatCo_3gbW4iB_rlSaaHVLAwKQUIY3hbOaC_6CwsI1PqhML90KBt_BbZnLT68c=)
* [Creating a Knowledge Graph with Neo4j: A Simple Machine Learning Approach](https://message.neo4j.com/NzEwLVJSQy0zMzUAAAF9pQWITzJyBAbRECrnLwJQk-yBYKsHMP2jAy3lq4CGTFxaYwaq37LtOG7l2IfBQcKLqd_Xjb0=)

(2) Run GraphQL to return listings that are in the best review score group giving a designated Price Group or Neighborhood.

# Results reporting:

August 3rd 2021 Presentation

August 9th 2021 Final Project

# References

Mark Needham & Amy Hodler (2019)

Graph Algorithms

Michele Castr（2021）

The Atlas for the Aspiring Network Scientist

# 

# Appendix

## Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Listing\_Detail Table** | | |
| **Field** | **Type** | **Description** |
| Id | integer | Airbnb's unique identifier for the listing |
| listing\_url | text |  |
| scrape\_id | bigint | Inside Airbnb "Scrape" this was part of |
| last\_scraped | datetime | UTC. The date and time this listing was "scraped". |
| name | text | Name of the listing |
| description | text | Detailed description of the listing |
| neighborhood\_overview | text | Host's description of the neighbourhood |
| picture\_url | text | URL to the Airbnb hosted regular sized image for the listing |
| host\_id | integer | Airbnb's unique identifier for the host/user |
| host\_url | text | The Airbnb page for the host |
| host\_name | text | Name of the host. Usually just the first name(s). |
| host\_since | date | The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest. |
| host\_location | text | The host's self reported location |
| host\_about | text | Description about the host |
| host\_response\_time |  |  |
| host\_response\_rate |  |  |
| host\_acceptance\_rate |  | That rate at which a host accepts booking requests. |
| host\_is\_superhost | boolean [t=true; f=false] |  |
| host\_thumbnail\_url | text |  |
| host\_picture\_url | text |  |
| host\_neighbourhood | text |  |
| host\_listings\_count | text | The number of listings the host has (per Airbnb calculations) |
| host\_total\_listings\_count | text | The number of listings the host has (per Airbnb calculations) |
| host\_verifications |  |  |
| host\_has\_profile\_pic | boolean [t=true; f=false] |  |
| host\_identity\_verified | boolean [t=true; f=false] |  |
| neighbourhood | text |  |
| neighbourhood\_cleansed | text | The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. |
| neighbourhood\_group\_cleansed | text | The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles. |
| latitude | numeric | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| longitude | numeric | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| property\_type | text | Self selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field |
| room\_type | text | [Entire home/apt|Private room|Shared room|Hotel]  All homes are grouped into the following three room types:  Entire place Private room Shared room Entire place Entire places are best if you're seeking a home away from home. With an entire place, you'll have the whole space to yourself. This usually includes a bedroom, a bathroom, a kitchen, and a separate, dedicated entrance. Hosts should note in the description if they'll be on the property or not (ex: "Host occupies first floor of the home"), and provide further details on the listing.  Private rooms Private rooms are great for when you prefer a little privacy, and still value a local connection. When you book a private room, you'll have your own private room for sleeping and may share some spaces with others. You might need to walk through indoor spaces that another host or guest may occupy to get to your room.  Shared rooms Shared rooms are for when you don't mind sharing a space with others. When you book a shared room, you'll be sleeping in a space that is shared with others and share the entire space with other people. Shared rooms are popular among flexible travelers looking for new friends and budget-friendly stays. |
| accommodates | integer | The maximum capacity of the listing |
| bathrooms | numeric | The number of bathrooms in the listing |
| bathrooms\_text | string | The number of bathrooms in the listing.  On the Airbnb web-site, the bathrooms field has evolved from a number to a textual description. For older scrapes, bathrooms is used. |
| bedrooms | integer | The number of bedrooms |
| beds | integer | The number of bed(s) |
| amenities | json |  |
| price | currency | daily price in local currency |
| minimum\_nights | integer | minimum number of night stay for the listing (calendar rules may be different) |
| maximum\_nights | integer | maximum number of night stay for the listing (calendar rules may be different) |
| minimum\_minimum\_nights | integer | the smallest minimum\_night value from the calender (looking 365 nights in the future) |
| maximum\_minimum\_nights | integer | the largest minimum\_night value from the calender (looking 365 nights in the future) |
| minimum\_maximum\_nights | integer | the smallest maximum\_night value from the calender (looking 365 nights in the future) |
| maximum\_maximum\_nights | integer | the largest maximum\_night value from the calender (looking 365 nights in the future) |
| minimum\_nights\_avg\_ntm | numeric | the average minimum\_night value from the calender (looking 365 nights in the future) |
| maximum\_nights\_avg\_ntm | numeric | the average maximum\_night value from the calender (looking 365 nights in the future) |
| calendar\_updated | date |  |
| has\_availability | boolean | [t=true; f=false] |
| availability\_30 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| availability\_60 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| availability\_90 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| availability\_365 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |
| calendar\_last\_scraped | date |  |
| number\_of\_reviews | integer | The number of reviews the listing has |
| number\_of\_reviews\_ltm | integer | The number of reviews the listing has (in the last 12 months) |
| number\_of\_reviews\_l30d | integer | The number of reviews the listing has (in the last 30 days) |
| first\_review | date | The date of the first/oldest review |
| last\_review | date | The date of the last/newest review |
| review\_scores\_rating | integer |  |
| review\_scores\_accuracy | integer |  |
| review\_scores\_cleanliness | integer |  |
| review\_scores\_checkin | integer |  |
| review\_scores\_communication | integer |  |
| review\_scores\_location | integer |  |
| review\_scores\_value | integer |  |
| license | text | The licence/permit/registration number |
| instant\_bookable | boolean | [t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing. |
| calculated\_host\_listings\_count | integer | The number of listings the host has in the current scrape, in the city/region geography. |
| calculated\_host\_listings\_count\_entire\_homes | integer | The number of Entire home/apt listings the host has in the current scrape, in the city/region geography |
| calculated\_host\_listings\_count\_private\_rooms | integer | The number of Private room listings the host has in the current scrape, in the city/region geography |
| calculated\_host\_listings\_count\_shared\_rooms | integer | The number of Shared room listings the host has in the current scrape, in the city/region geography |
| reviews\_per\_month | numeric | The number of reviews the listing has over the lifetime of the listing |

|  |  |  |
| --- | --- | --- |
| **Listings Table** | | |
| **Field** | **Type** | **Description** |
| Id | integer | Airbnb's unique identifier for the listing |
| name | text | Name of the listing |
| host\_id | integer | Airbnb's unique identifier for the host/user |
| host\_name | text | Name of the host. Usually just the first name(s). |
| Neighbourhood\_group |  |  |
| neighbourhood | text |  |
| latitude | numeric | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| longitude | numeric | Uses the World Geodetic System (WGS84) projection for latitude and longitude. |
| room\_type | text | [Entire home/apt|Private room|Shared room|Hotel]  All homes are grouped into the following three room types:  Entire place Private room Shared room Entire place Entire places are best if you're seeking a home away from home. With an entire place, you'll have the whole space to yourself. This usually includes a bedroom, a bathroom, a kitchen, and a separate, dedicated entrance. Hosts should note in the description if they'll be on the property or not (ex: "Host occupies first floor of the home"), and provide further details on the listing.  Private rooms Private rooms are great for when you prefer a little privacy, and still value a local connection. When you book a private room, you'll have your own private room for sleeping and may share some spaces with others. You might need to walk through indoor spaces that another host or guest may occupy to get to your room.  Shared rooms Shared rooms are for when you don't mind sharing a space with others. When you book a shared room, you'll be sleeping in a space that is shared with others and share the entire space with other people. Shared rooms are popular among flexible travelers looking for new friends and budget-friendly stays. |
| price | currency | daily price in local currency |
| minimum\_nights | integer | minimum number of night stay for the listing (calendar rules may be different) |
| number\_of\_reviews | integer | The number of reviews the listing has |
| last\_review | date | The date of the last.newest review |
| reviews\_per\_month | numeric | the number of reviews the listing has over the lift time of the listing |
| calculated\_host\_listings\_count | integer | The total numbers of host listings count |
| availability\_365 | integer | avaliability\_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host. |