Assessment 1 - GEOGM0068: Geographic Information Retrieval and Integration

Database for predicting Airbnb house prices

student name: Jia Zhao

student id: 2472581

word count: 1864

1. Introduction

1.1 Why Choose Airbnb Dataset

The Airbnb dataset utilized in this report comprises three tables: listings_d, which records property information; reviews_d, detailing each review's content and timestamp; and calendar_d, documenting rental prices for properties at different times.

The Airbnb dataset boasts rich data dimensions, with a particularly intriguing focus on:

- Starting from property IDs, each ID is associated with comprehensive rental information, guest ratings, host descriptions, spatial locations, and guest reviews. Despite numerous tables and dimensions, the dataset maintains a high degree of organization.
- It contains abundant time series information, enabling exploration of property similarity, consistency, and potential patterns over time and space.
- The diverse dimensions all converge on a common theme: the guest experience. The emergence of various dimensions aims to provide guests with a comprehensive understanding of properties, aid in their stay, and enable post-order evaluations. This symbiotic service relationship also represents the interconnectedness among the dataset's various dimensions.

These attributes render the Airbnb dataset highly conducive to research, whether conducted in isolation or through the integration of multiple datasets.

1.2 Topics that can be studied with this dataset

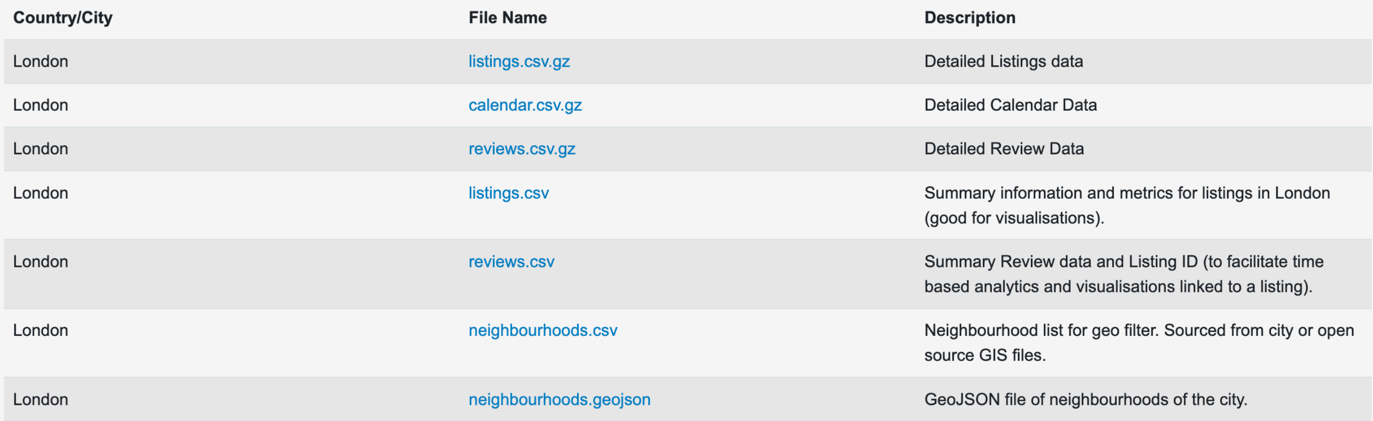
This dataset holds significant research value for various societal stakeholders. It clearly delineates its primary participants: hosts and guests. These two entities alone offer a plethora of research topics, such as how guests select suitable accommodations and how hosts manage their assets or set prices.

However, if we shift our focus from this basic supply-demand relationship, we can examine relationships beyond hosts and guests to include hosts and society, hosts and government, Airbnb and society, and Airbnb and government. This broader perspective allows for the exploration of issues related to gentrification, community displacement, and cultural erosion within the real estate domain. For example:

- Cox. (2018)explores the likelihood of Airbnb hosts in black communities in New York City being three times more likely to be white than black. Additionally, among white hosts, only a small fraction reside in black neighborhoods, which disrupts these communities.
- Gutiérrez et al. (2016)contrast the sudden emergence of the Airbnb model with the traditional hotel model in the tourism industry in Barcelona.
- Guttentag et al. (2017)employ cluster analysis to identify five motivating factors influencing guests' choice of accommodations: interactivity, family amenities, novelty, the spirit of the sharing economy, and local authenticity.

1.3 Dataset acquisition source

- Inside Airbnb(http://insideairbnb.com/)
  Inside Airbnb is a mission driven project that provides data and advocacy about Airbnb's impact on residential communities. They work towards a vision where data and information empower communities to understand, decide and control the role of renting residential homes to tourists.



Country/City	File Name	Description
London	listings.csv.gz	Detailed Listings data
London	calendar.csv.gz	Detailed Calendar Data
London	reviews.csv.gz	Detailed Review Data
London	listings.csv	Summary information and metrics for listings in London (good for visualisations).
London	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
London	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
London	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

Figure 1 Data provided by Inside Airbnb

- OpenStreetMap（https://www.openstreetmap.org/)
  Because city-level POIs are difficult to obtain, I used the API to download them locally for further processing.

2. Method

2.1. Basic information of the data table

The current report utilizes Airbnb data from three cities spanning the years 2010 to 2024. Each city dataset comprises three tables:

- `listings_d` : This table contains detailed attributes of the properties (such as spatial coordinates, overall ratings, room details, basic descriptions) as well as partial information about the hosts (such as host ID, number of listings, response rate).

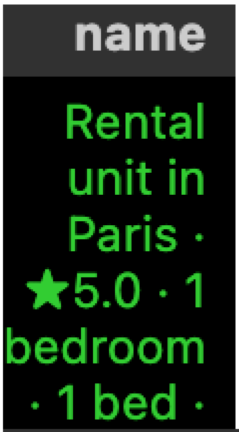
- `reviews_d` : This table records all reviews for the properties during the study period.
- `calendar_d` : This table captures time-series information regarding property availability and price trends during the study period.

All three tables share a common field, "listing_id," which serves as the unique identifier for each property and acts as the carrier for various property attributes and records.

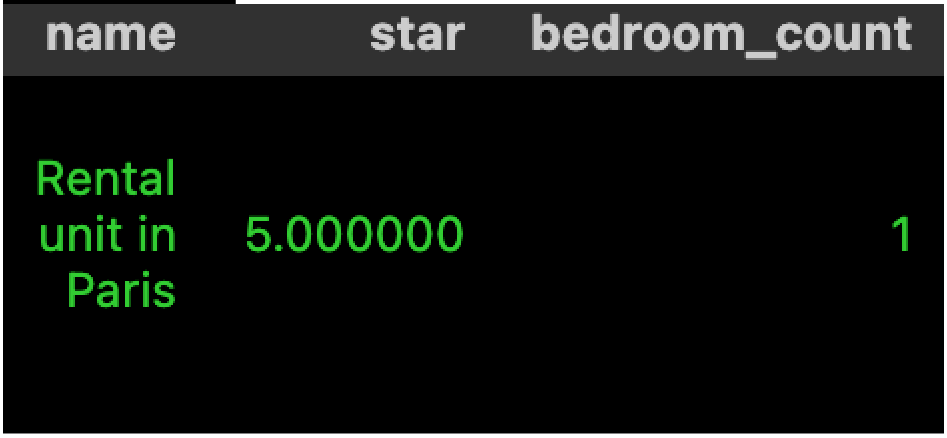
Additionally, the dataset includes Points of Interest (POIs) data, which documents spatial point information such as public transportation, public services, and open spaces within the cities.

2.2 Data preprocessing

- To streamline the `listings_d` table for this report, we will remove redundant, irrelevant, and overly detailed fields, reducing the total number of fields from approximately 50 to 18. Additionally, certain fields, such as `calendar_last_scraped` require standardization to extract relevant information. Here are the steps:
  - Remove fields with empty or irrelevant content.
  - Eliminate redundant fields or those duplicating information.
  - Retain essential fields relevant to the report's focus.
  - Standardize fields requiring further processing, such as "bathroom_text."
    After these steps, we'll have a more concise and relevant dataset for further preprocessing.
- The contents of some fields need to be standardized. For example, the `bathroom_text` column contains information such as `2 bathrooms` and `1 shared bathroom` , and we need to further split it.
- Some fields contain multiple meanings and need to be split, as shown below.



name
Rental unit in Paris · ★5.0 · 1 bedroom · 1 bed ·



name	star	bedroom_count
Rental unit in Paris	5.000000	1

Figure 2 The original name field (left), the field after splitting as needed (right)

- Attribute correction: The types of many fields in the table are wrong, which will cause subsequent retrieval failures, so the fields need to be converted into the types we need.
  For example: convert `Date` Field from `text` to `date` type, and some fields from `text` to `num` type.
- Some unfinished data preprocessing
  NLP: This data set contains a total of approximately 4.5 million comments(Took two hours to figure out non-English comments). Whether batch processing or parallel processing requires a lot of time, and data storage is also time-consuming and laborious.Hong et al. (2015) mentioned that text features can be extracted in the corpus matrix and cluster analysis can be performed on the features, which can avoid traversing each review.

2.3 Data table connection

- Airbnb data set relationship definition:
  - `listings_d` and `reviews_d` : A listing can have multiple reviews, but a review only belongs to one listing (one-to-many relationship)
  - `listings_d` and `calendar_d` : one listing corresponds to availability and price information on multiple dates (one-to-many relationship)
- Specific methods
  - Make sure the data types of the two columns are consistent, such as integers or strings.
  - Since the id of the `listings_d` table is unique, and the `listing_id` of the `reviews_d` table may be repeated, you can check whether each id in `listings_d` appears at least once in the listing_id of `reviews_d` .
  - Perform a correlation check to see if `reviews_d` and `listings_d` can be connected by `listing_id` and id without error. This can also help confirm the correspondence between the two.
  - Randomly select some `listing_ids` , check whether they exist in the id column in the `listings_d` table, and whether the related information matches.
- Link to geospatial database: Connect the Qgis project to postgregis, let the `POIs data set` and `Airbnb data set` be under one database, and install the `PostGIS` extension for subsequent analysis.
- Index optimization: Create indexes for commonly used query columns (such as `listing_id`) to improve query efficiency.

3. Function

3.1 Basic search

Now it can do some preliminary queries, for example:

- In order to determine the occupancy status of properties with different star ratings in a certain period, we can query: How many reviews did properties with 4.8 stars and 4.7 stars or above have in July 2020?

City	4.8 Stars Reviews	4.7 Stars Reviews
NY	1684	2329
LD	1255	1839
PR	2546	4004

Table 1 The total number of reviews of properties with 4.8 stars and 4.7 stars or above in each city in July 2020

- This database also can be used to assess whether the presence of bus stations affects property prices. By utilizing geographic spatial database extensions and setting search ranges to 500m and 1000m, we can retrieve properties within these buffer zones that lack bus stations and compare their average prices with those of properties located near bus stations. The result:
  - with a transportation stop within 500 meters is 195.18
  - with transportation stops within 1,000 meters is 193.74
  - without transportation within 500 meters is 186.08
  - without transportation within 1000 meters is 162.11

It can be seen that bus stops are indeed an influencing factor on housing prices, and there is a gap in the prices of houses without bus stops within one thousand meters.

3.2 House price forecast

We take London as an example and conduct a simple prediction analysis of Airbnb house prices in London through KNN and property characteristics. The main selected independent variables are as follows:

- Whether it is the entire property
- Is there a separate toilet?
- Property rating
- Total number of reviews for the property
- Number of public services, transportation, and open space POIs near the house, including development spaces (historic buildings, universities, shopping malls, playgrounds, parks, etc.)

Perez-Sanchez, Serrano-Estrada, Marti, and Mora-Garcia (2018)For details on the study of the correlation between property characteristics and property prices, see Appendix. The prediction steps are as follows:

- Link the Airbnb listings dataset with the London Points of Interest (POIs) dataset. The Airbnb dataset contains latitude and longitude information as coordinates, while the London POIs dataset is in shapefile (shp) format and does not include latitude and longitude information. Instead, it has a "geom" field, requiring conversion using the PostGIS extension "geography."

- Divide the dataset into training and testing sets for `KNN`.
  The reliability of KNN house price prediction in this database is as follows:
  - Best parameters: {'n_neighbors': 30}
  - Best score (negative MSE): -90048.73146182715
  - MAE: 105.47686022205934
  - MSE: 270772.86292522924
  - R^2 Score: -0.012885747175823248

Judging from the prediction results, the error is very large, and the value of the optimal n_neighbors is also abnormally large. This may be mainly attributed to the oversimplified prediction process and the missing feature engineering link.

4. Discussion

4.1 Advantages

- The data structure is clear and can be further sorted out
  - Architectural attributes and geographical attributes of the property
  - Social and economic attributes of the landlord
  - Passengers' travel purposes and housing needs
- Extensible, the data preprocessing process and data table links are reusable
  - Laterally, you can continue to add Airbnb data sets from different cities and different landforms.
  - You can add Airbnb data sets at different times vertically
  - Because there are databases with spatial attributes, they are also compatible with diverse geographic data sets.
- This data set can be used for various research topics, not just for studying Airbnb housing prices or using Airbnb as the research subject. Hosts, travelers, and communities can all be used as research objects. At the same time, the applicability of the research field is also wide.

4.2. Limitations

- Retrieval efficiency is relatively low
  - Because it is necessary to search from a large number of spatial objects
  - Each spatial object also carries a lot of information. A house can be regarded as a small database.
  - This database does not consider vectors or more efficient data storage modes
- Data sources are all from metropolitan areas, and the diversity is poor. Richer regional data need to be collected, such as non-urban areas.

- The amount of comment text data is too large. It took more than 40 minutes to check whether the comments were in English. Perhaps a more complete database requires the support of SparkSQL.

References

- Cox. (2018). A Year Later: Airbnb as a Racial Gentrification Tool. http://insideairbnb.com/research/a-year-later-airbnb-as-a-racial-gentrification-tool/
- Gutiérrez, J., García-Palomares, J., Romanillos, G., & Salas-Olmedo, M. (2016). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. Tourism Management, 62, 278-291. https://www.sciencedirect.com/science/article/abs/pii/S0261517717301036
- Guttentag, D., Smith, S., […] & Havitz, M. (2017). Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. Journal of Travel Research, 57(3). https://journals.sagepub.com/doi/full/10.1177/0047287517696980?casa_token=__dY1B39mX4AAAAA%3AxjbAR8E2R6iUJoGO7fhs2SmjF-2OW8oR-c02-f2SFH16Pie66_xr-rAghMUeagqRzCCM2yLSqvgn4g
- Hong, S.-S., Lee, W., & Han, M.-M. (2015). The Feature Selection Method based on Genetic Algorithm for Efficiency of Text Clustering and Text Classification. International Journal of Advanced Soft Computing and Applications, 7(1), March. ISSN 2074-8523.https://www.i-csrs.org/Volumes/ijasca/3.Sung-Sam-Hong-et-al.pdf
- Perez-Sanchez, V.R., Serrano-Estrada, L., Marti, P., & Mora-Garcia, R.-T. (2018). The What, Where, and Why of Airbnb Price Determinants. Sustainability, 10(12), 4596. https://www.mdpi.com/2071-1050/10/12/4596

Appendix：

Accommodation characteristics	Coefficient	Ad and host features	Coefficient
`listing_type`	0.563	`ad_online_duration`	0.013
`bathrooms`	0.157	`rating`	0.024
`max_guests`	0.061		
`secu_deposit`	0.0002		
`cleaning_fee`	0.005		
`cancella_policy`	0.032		
`num_photos`	0.001		

Environmental characteristics	Coefficient	Location characteristics	Coefficient
`cat_sightseeing`	0.152	`alicante`	Reference
`cat_eating`	0.054	`valencia`	0.165
`cat_shopping`	0.048	`con_coastalfringe`	0.109
		`discon_coastalfringe`	-0.115
		`coastal_dist_km`	-0.027
		`inte_cat_dist_km`	0.013

Table 2: Factors Affecting Property Prices

Note: The code in the attachment is only the Python part. Because most operations are completed in SQL, the attachment does not show all the experimental processes.