AirBnB: Relational Ratings

Jhernie Evangelista, Georgia Myers, Lija Hoffman, Niurika Gonzalez

Summary:

- focused on exploring AirBnB listings in Chicago, IL
- utilized a RapidAPI and a Kaggle dataset in .csv form
- Cleaned and reshaped data for a clearer merge
- Casted the final data into a relational database

Extract

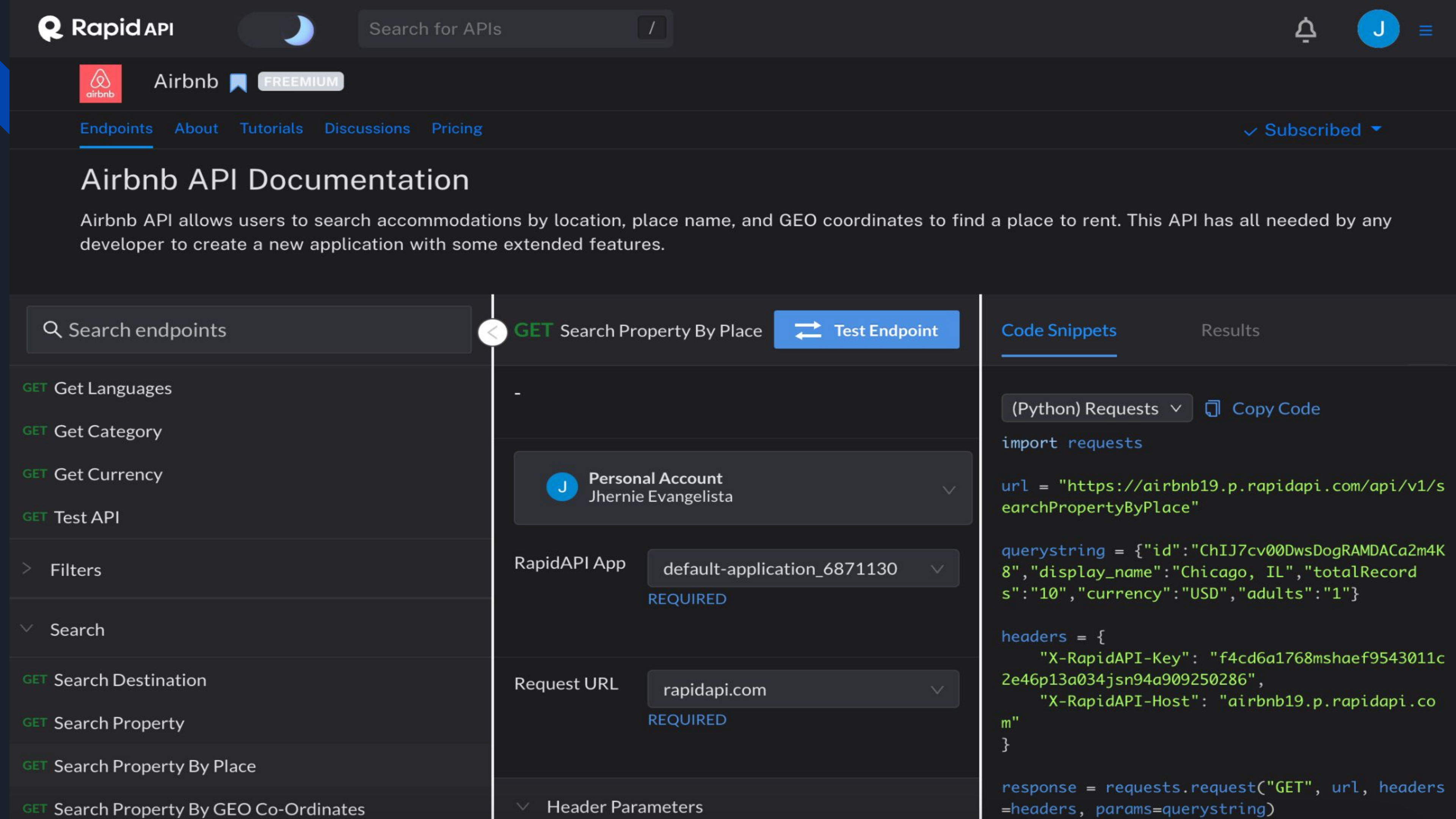
1. Queried the RapidAPI source using [Search Property by Place end point.]

    *Did this x5 due to fixed guest count per query

2. The Kaggle dataset had AirB&B listings in .csv form for all major US Cities

3. Using these, we identified fields unique to each dataset, justifying the need to identify more than one datasource. Some unique fields are listed below:

    * "Rating"
    * "Adults"
    * "Baths"
    * "Beds"



Airbnb API Documentation

Airbnb API allows users to search accommodations by location, place name, and GEO coordinates to find a place to rent. This API has all needed by any developer to create a new application with some extended features.

GET Search Property By Place

-

Personal Account
Jhernie Evangelista

RapidAPI App: default-application_6871130
REQUIRED

Request URL: rapidapi.com
REQUIRED

```python
import requests

url = "https://airbnb19.p.rapidapi.com/api/v1/searchPropertyByPlace"

querystring = {"id":"ChIJ7cv00DwsDogRAMDACa2m4K8","display_name":"Chicago, IL","totalRecords":"10","currency":"USD","adults":"1"}

headers = {
    "X-RapidAPI-Key": "f4cd6a1768mshaef9543011c2e46p13a034jsn94a909250286",
    "X-RapidAPI-Host": "airbnb19.p.rapidapi.com"
}

response = requests.request("GET", url, headers=headers, params=querystring)
```

Transform

* Initially we wanted to use the PostgreSQL tool but struggled to import the .csv file into PGAdmin.

*possibly due to a number of special characters used in the "listing name" field

* We resorted to using Pandas dataframes by pulling the data into a jupyter notebook file via ".read_csv" after manually removing unwanted columns such as "neighbourhood-group"

* Used a "for loop" to extract the relevant datapoints from the Air B&B api.

* Used Pandas .concat method on the API queries to create one main table for all of the pulled data.

Transform cont’d.

* To merge:

1. Identified the common field between both datasets— "unique listing ID."

2. Modify the fields to use the same datatype

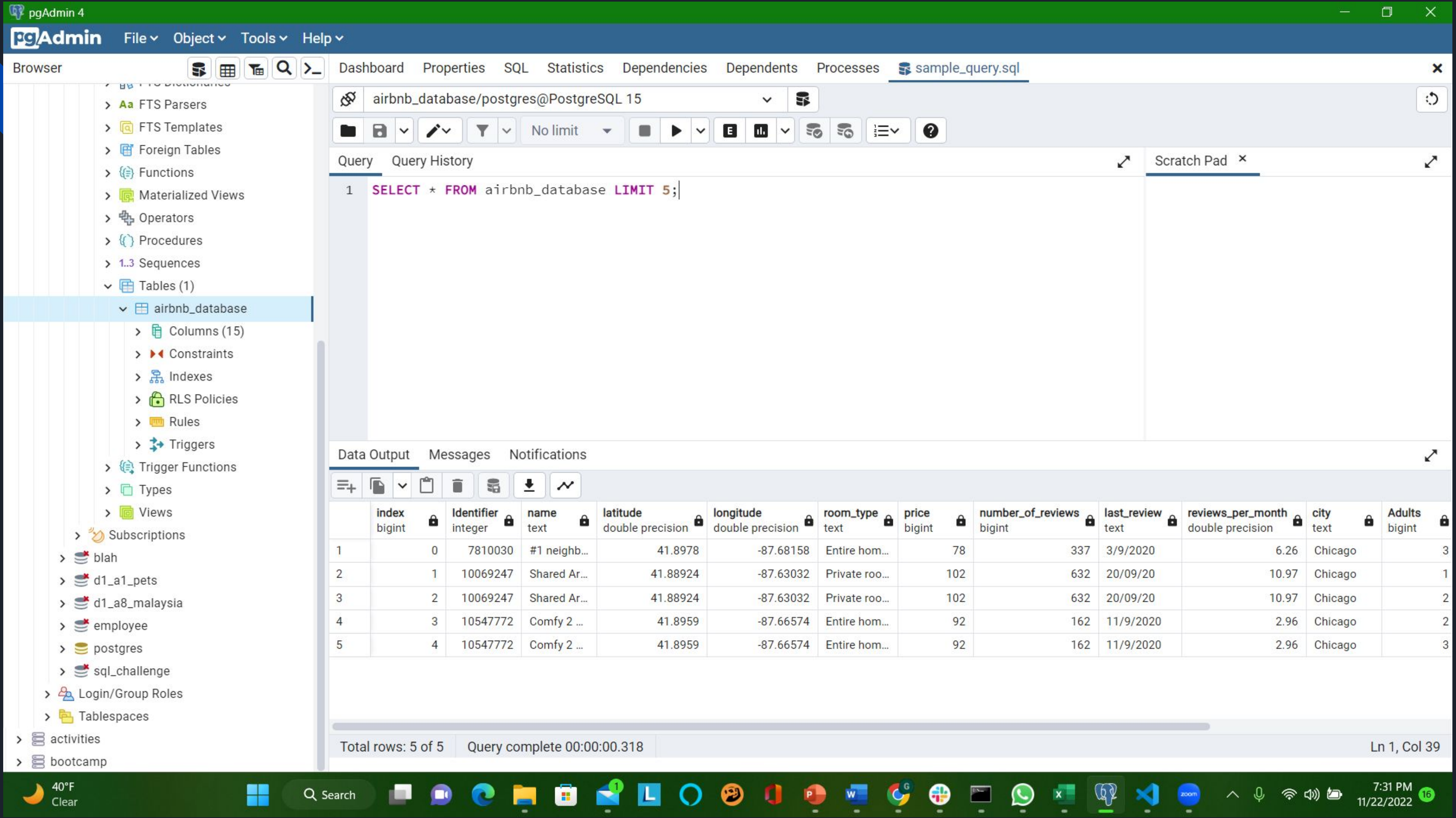
    *one set’s field was an "object" and the other’s an "int."

        -fixed using .astype(np.int64)

3.Used an inner merge to combine the cleaned datasets into one that held only data relating to Chicago listings.

Load

* We created a new database using PostgreSQL in the pgAdmin tool
* Load the main table into the database using SQLAlchemy
* Push the dataframe into the database using the built connection
* Once connected, a user can perform queries on the table using SQL, such as the query depicted



airbnb_database/postgres@PostgreSQL 15

```sql
SELECT * FROM airbnb_database LIMIT 5;
```

	index bigint	Identifier integer	name text	latitude double precision	longitude double precision	room_type text	price bigint	number_of_reviews bigint	last_review text	reviews_per_month double precision	city text	Adults bigint
1	0	7810030	#1 neighb...	41.8978	-87.68158	Entire hom...	78	337	3/9/2020	6.26	Chicago	3
2	1	10069247	Shared Ar...	41.88924	-87.63032	Private roo...	102	632	20/09/20	10.97	Chicago	1
3	2	10069247	Shared Ar...	41.88924	-87.63032	Private roo...	102	632	20/09/20	10.97	Chicago	2
4	3	10547772	Comfy 2 ...	41.8959	-87.66574	Entire hom...	92	162	11/9/2020	2.96	Chicago	2
5	4	10547772	Comfy 2 ...	41.8959	-87.66574	Entire hom...	92	162	11/9/2020	2.96	Chicago	3

Total rows: 5 of 5   Query complete 00:00:00.318   Ln 1, Col 39

Considerations/Reflections

* During project worktime, we realized that joining the data from the API pulls could have been done within the database, which would have enabled queries on already-filtered data.

* RapidAPI is a tool that anyone can contribute to, so the validity of the pulled data is questionable.

* With more time, we would have web-scraped the AirBnb platform to gain missing data and add it as a table into the database.

Potential Analysis

Analysis could be performed on a number of metrics included in the tables procured.

Examples include:

* Is there a trend amongst listing titles and rate of occupancy per listing or overall rating?

* Does the guest count cap affect the rate of bookings per month for each listing?

* With additional datasources such as median home prices, could also analyze the relationship between concentration of AirBnB listings in a particular geography, and the average median home price/value.

Questions?