WanderJaunt

Data Science Challenge

Data on short term rental prices and occupancy is very important to WanderJaunt. It helps inform us how competitors are pricing, which influences our own pricing strategy, and helps us benchmark our own occupancy and revenue per available room against similar properties. In addition, it provides key inputs to the decision of what locations and markets we enter and what types of properties can be the most profitable.

Clean data on performance in the short-term rental market can be hard to obtain, but signals can be derived from the public posting of price and availability on public platforms like Airbnb. Though messy, scraping this data is possible. The dataset provided is the daily scraped availability and price for Airbnb listings in the Phoenix market for all dates from 4/1/18 to 5/31/18. The two data tables provided as flat CSV files are structured as follows:

scraped_listings.csv
This table has 1 row for each `scraping_id` (rentable unit at a property) with various information about that listing. The listings have been filtered to include only units that are entire homes / apartments for rent (not individual / shared rooms). See below details on the field in this table:

scraping_id – ID key for the unit that joins to the scraped_data.csv table
listing – URL link to the Airbnb posting
city – Name of city within the Phoenix market
lon – Longitude of the unit rented
lat – Latitude of the unit rented
mapped_location – A google maps URL of the unit's location
name – Posting name
capacity – Number of people unit is said to accommodate
bathrooms – Number of bathrooms at unit
bedrooms – Number of bedrooms at unit
has_pool – 1 if unit has pool listed as amenity; 0 if does not
cleaning_fee – The amount in dollars charged at checkout to cover cleaning cost
is_superhost – 1 if the host account is a superhost; 0 if not (https://www.airbnb.com/superhost)
host_name – Name of host

scraped_data.csv
This table has the availability and price of all listings for April to May 2018. This data is scraped daily and the results for each day between 3/15/18 and 5/31/18 are provided in this table. See below details on fields:

scraping_id – ID key for the unit that joins to the scraped_listing.csv table
as_of_date – Date the information was scraped
date – Date of the night to be booked
price – Price in dollars of the night (excluding taxes/fees/cleaning)
available – 1 if unit is available to be booked; 0 if not

Notes
The scraped data is messy and available = 0 may not necessarily mean the unit was booked and generated revenue. Hosts may block dates or ranges of dates as unavailable. You may want to consider building logic to exclude what you suspect to be a block from revenue and occupancy calculations.

Questions to answer

- What data would you exclude from analysis for being unreliable or potentially a block instead of an actual booking?
- What is a good approach to estimate occupancy and revenue per unit?
- In which month do properties appear to generate more revenue? April or May?
- How much more revenue do places with 3 bedrooms make vs. places with 2 bedrooms?
- What are any other interesting insights you may have found?