Inside Airbnb's data on Austin TX

Explanation: why this dataset?

Convergence of a few of my key interests:

- Real estate/ short term rental (STR) market
  - I currently manage a unit on Airbnb/VRBO.
- Austin, TX
  - Lived there for a couple years, have friends there.
  - Experienced rapid growth which has slowed…but to what extent?
    - Very much buoyed by tech.
- Data Analysis!
  - Need a final project with an attractive Tableau dashboard – holds the potential for any number of helpful visualizations that communicate:
    - what makes a successful STR in the Austin area
    - neighborhood by neighborhood visualizations
    - time series data from previous scrapes/data iterations.
      - Not hand-hold level; will have to figure out how to achieve this.
  - Data Requirements:
    - Open Source – check
    - Authentic/Authoritative source – Inside Airbnb credible
      - Stance/angle is monitoring effects on housing markets. Watchdog.
    - Non-anonymized column names – checked in Jupyter.
    - No more than 3 years old – it's barely three months old.
    - At least 2 continuous variables
      - Price, review scores (and subcategories), reviews per month
        - Need to investigate price – dynamic pricing is common practice. How does that translate to the scrape?
          - AirDNA cross-reference by listing ID?
        - Review Score is an aggregate of these sub-categories:
          - Review_scores_accuracy
          - Review_scores_cleanliness
          - Review_scores_checkin
          - Review_scores_communication
          - Review_scores_location
          - Review_scores_value
    - At least 2 categorical variables

            - Hosts: Pros vs Regular Joes
                - Would need to derive, should be easy.
                - Host_listings_count vs host_total_listings_count?
            - Superhost t/f
            - Property_type and room_type
            - Neighborhoods/zip codes
            - Could derive price ranges
            - Instantly bookable t/f
        - Contain at least 1,500 rows
            - Well over 1.5K, the listings df alone was 15K rows.
        - Geographical component
            - Lat/longs
            - Zip codes
        - Time series
            - Potentially. Could grab older versions of the data and try to create time series.

Data Source

- Internal or External? Mainly internal but kind of both.
    - Scraped from Airbnb's listings, which is public information.
    - There are assumptions taken:
        - Occupancy rate is estimated based off of review rate.
            - Estimated by avg. length of stay x estimated bookings.

Data Collection Methods

- Scraped from listing data which is Administrative.
- Reviews data would be more like Interviews and Surveys
    - Collection bias, sometimes only really good or really bad stays get reviewed at all.

Data Contents

- **Listings** dataset the meatiest, will be the backbone of my analysis.
- Inside Airbnb includes helpful data dictionary: https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596
- Property type could be interesting category (guesthouse, bungalow, private room, etc.)
- Reviews per month an interesting metric.

Data Relevance

- Highly relevant data by my estimation. Very current with DEC 2023 data, and with the option to grab previous quarterly data for time series analysis.
    - Would have to request archived data for anything beyond the last 4 quarters, which might be worth it in order to get YoY trend analysis.
        - Particularly interesting because of Austin TX designation as a tech boom town, especially preceding and during COVID; they've since experienced a slight market correction. How, if at all, has this affected the STR market?
        - Basic but key Business question: how many Airbnb's are in the Austin market, and how has that number changed over time?
    - Need to take care to adequately preface the assumptions taken for estimated bookings and occupancy rate – they are not firm figures disclosed by Airbnb but are rather deductions based on prevailing statistics and public facing data.

Data Profile

Data Cleaning

- Heads are not anonymized which is good.
- The first thing that I'm going to do is limit the listings df columns and only include what I need.
    - Keepers include:
        - id, listing_url, host_id, host_since, host_response_rate, host_acceptance_rate, host_is_superhost, host_neighbourhood, host_listings_count, host_verifications, host_has_profile_pic, host_identity_verified, neighbourhood_cleansed, latitude, longitude,, property_type, room_type, accommodates, bathrooms_text, bedrooms, beds, amenities, price, minimum_nights, maximum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, has_availability, availability_30, availability_60, availability_90, availability_365, number_of_reviews, number_of_reviews_ltm, number_of_reviews_l30d, first_review, last_review, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, instant_bookable, calculated_host_listings_count, calculated_host_listings_count_entire_homes,

calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, reviews_per_month

- amenities didn't look like it came thru though, should be an array
- consider how much utility the max_min, min_min, max_max, and min_max columns have
- consider utility/uses for the availability variables (has availability, availability_30, availability_60, availability_90, availability_365)
    - Could categorize listings/owners by how far in advance they probably set their calendars.
    - Key dataset blind spot: Doesn't distinguish between a booking and a block.
- Can estimate listing age by first_review
    - Not definitive though; listings can be duplicated.
- 53 columns

- Maybes include:
    - source, name, neighborhood overview, host_about, bathrooms (Showing all NaN, bathrooms text is actual – same for bedrooms)
        - going to need to cross-reference some listing id's with the listings on site, to see if bedrooms and beds are correlated at all.

Understanding your data

- Created copy of data dictionary in Google Sheets
    - Can track which columns were phased out and at what df iteration.
    - Copied .describe() outputs and taking notes – pseudo data profile without getting too granular on 50ish variables.
        - Basic statistical analysis from .describe() function:
            - Count, mean, standard deviation, min, 25%, 50%,75% and max.

Considering Limitations and Ethics

Limitations

- Data scraped from public-facing listings as opposed to internal Airbnb data so:
    - Occupancy is an *estimate*.
        - Review Rate of 50% used to convert reviews to estimated bookings. See methodology here:

                - http://insideairbnb.com/data-assumptions/
            - Average length of stay, where available, is configured for each city, multiplied by the estimated bookings for each listing over a period gives the occupancy rate.
                - When no public statement is made about average stay, a value of 3 nights per booking was used.
                - If a listing has a higher minimum nights value than the average length of stay, the minimum nights value was used instead.
            - Austin appears to be using the 3 nights per booking rule of thumb, with the review rate of 50% to convert reviews to estimate bookings.
                - In short, a unit with 21 reviews in a year and a one night minimum is estimated to have been occupied for 126 nights. Whereas a room with 3 reviews and a 30 night minimum is estimated to have been occupied for 180 nights.
                    - 21*3 = 63 / .5 = 126
                    - 3*30=90 /.5 = 180
    - Listing scrapes are a snapshot.
        - Pricing ***can be changed*** on any unoccupied date. Many professional hosts will utilize dynamic pricing to move unsold nights, which we're unable to account for at least with one scrape of data.
            - Will compare with previous scrape dates for time series.
        - Blackout dates can and do change frequently. Additionally hosts choose how far out in advance to open up their calendars, from only 1 month out to 6 months or a year out.
            - Availability variables can shine some light on how far out hosts are choosing to list in advance.
        - Listings can be deleted.
            - Often with a duplicate of the listing, can wipe a slate clean of any reviews (negative or positive)
    - Review Scores have collection bias.
        - Review rate not uniform; people may only rate when experience is very positive or negative.
            - Will probably lower the estimates for mediocre ratings.
        - Additionally, reviews are inherently biased at least on a case by case basis.
            - Hopefully with more ratings comes more reliability.
- When comparing profitability of these units, we should take them with a grain of salt because they are a compound of estimates.
    - Pricing (which can change) x Occupancy (review rate /.5 x 3 (or min nights))

Ethics

- No private information is being used. Names, photographs, listings, and review details are all publicly explained on Airbnb website.
- Location information for listings is anonymized by Airbnb.
    - In practice, this means the location for a listing on the map or in the data will be 0-450ft of the actual address.
    - Listings in the same building are anonymized by Airbnb individually, and therefore may appear "scattered" in the area surrounding the actual address.
- Inside Airbnb's mission is to track the impact of Airbnb's impact on residential areas in which hosts are renting out residential properties as hotels, as opposed to sharing the primary residence in which they live occasionally.
    - Though serving as a watchdog organization of sorts, they are aggregating publicly available information for which to illuminate a morally gray situation.
        - This site claims "fair use" of any information compiled in producing a non-commercial derivation to allow public analysis, discussion and community benefit.
    - My analysis may in some sense contradict this mission; most of its insights will be tailored toward prospective hosts, investors, and tourists.
        - Mostly it will be to showcase my skills as an Analyst.3
        - Would appreciate feedback on which of these to tailor/prefer for my audience.

Questions to Explore

- What's the most important factor to be a successful listing on Airbnb?
- What proportion of hosts in the Austin area are professional hosts?
    - What's the cutoff on listings for professional hosts?
    - Should we compare single listing hosts vs 10+ listing hosts, or is this almost a different category?
- What are the most popular neighborhoods in Austin for Airbnb?
    - Most profitable?
    - Highest rental rate?
    - Most rentals?
    - Busiest outright?
        - Any correlation with walk-score?
- What are the most in-demand amenities for rentals in Airbnb?
    - Would need to fix amenities variable, which appears to be json data type?
- Incumbents vs Challengers: do older/more established listings outcompete newer ones?
    - If so, to what extent?
    - When does a new listing cross over into established territory.

- Are Private or Shared rooms a factor at all in the Austin market?
    - How big a portion of all rentals are they?
    - How much do they go for?
    - In what parts of town are they viable, versus whole properties?
- Comparing self-selected property types: Which is the most expensive? The most popular?
- What months are the busiest/most expensive for Austin Airbnb?
    - Pricing
    - Availability
- How do review scores correlate to price? Do some review scores (accuracy, value) have a stronger correlation than the others?
- Impact of the… (Booleans)
    - Superhost badge
    - Instant book
    - Host identity verified
    - Profile Pic
- What's the coolest neighborhood in Austin?
- Implementing Customer Profiles
    - The cashflow investor
    - The vibes vacationer/ The traveling artist
    - The prospective roomie renter

Cleaning today

- Create sub-dfs for shared and entire homes
    - If not substantial shared homes (1% of total listings) sequester it from further analysis
- Figure out bathroom variable, consolidate with text variable.