Project Information

Austin Tam

European Airbnb Dataset

Data Source

This dataset contains data about Airbnb rental locations across multiple cities in Europe. It was composed for an academic article that can be found here. However, the data itself can be found here. Given that it was used for an academically rigorous article, we have reason to trust the data source.

Data Contents

The dataset contains a multitude of variables for each Airbnb listing across the following European cities: Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, London, Paris, Rome, Vienna. The columns of the dataset are the following:

Variable	Description
realSum	the full price of accommodation for two people and two nights in EUR
room_type	the type of the accommodation
room_shared	dummy variable for shared rooms
room_private	dummy variable for private rooms
person_capacity	the maximum number of guests
host_is_superhost	dummy variable for superhost status
multi	dummy variable if the listing belongs to hosts with 2-4 offers
biz	dummy variable if the listing belongs to hosts with more than 4 offers

cleanliness_rating	cleanliness rating
guest_satisfaction_overall	overall rating of the listing
bedrooms	number of bedrooms (0 for studios)
dist	distance from city centre in km
metro_dist	distance from nearest metro station in km
attr_index	attraction index of the listing location
attr_index_norm	normalised attraction index (0-100)
rest_index	restaurant index of the listing location
attr_index_norm	normalised restaurant index (0-100)
lng	longitude of the listing location
lat	latitude of the listing location

Data Limitations

- The data looks like it is only obtained at one point in time, meaning that it cannot capture the seasonality of Airbnb prices (e.g. during popular times of travel, like Christmas, prices will be higher).
- If the traveler books the Airbnb within 4 months of travel, the user will experience Airbnb's smart pricing. Which means that each user will see a different price. Therefore, we must take this limitation into account when assessing the realSum column in the data frame.
- There may be some omitted variable bias due to the limited nature of the dataset. For example, though hard to quantify, the dataset could have accounted for the amenities that the Airbnb has, as this would surely affect the prices of the Airbnb.

Data Ethics

The main concern in terms of data ethics with this dataset are the lat and long of each airbnb listing in the dataset. Usually on the Airbnb website, each listing only has a general

location until the user books the Airbnb and then its precise location is revealed. Therefore, it may be problematic that this dataset contains precise latitude and longitude data of each listing. This issue may be lessened by the fact that there is not any other identifying information for the Airbnb listing that is contained within the dataset.

Eurostat Temporal Airbnb Dataset

Data Source

The dataset contains data about Airbnb vacation stays throughout Europe from 2018 to 2022. The European Union worked alongside Airbnb to provide this data, thus we have ample reason to trust this dataset. The data can be found here.

Data Contents

The dataset contains a multitude of variables for each Airbnb listing across the following European cities: Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, London, Paris, Rome, Vienna. The columns of the dataset are the following:

Variable	**Description**
indic_to,c_resid,month,unit,geo\time	Various variables combined into one. The documentation was not very clear, but for our purposes we just needed the month, and geo/time variables which gave us the month and country of the airbnb stay.
2018	Contains the number of stays for year 2018
2019	Contains the number of stays for year 2019
2020	Contains the number of stays for year 2020
2021	Contains the number of stays for year 2021
2022	Contains the number of stays for year 2022

Data Limitations

- The data is only at a country level which means that we cannot conduct an analysis on a city level. However, since we are trying to capture the seasonality of travel for these countries, it should not be too big of a deal
- There are no metrics on how long people are staying at Airbnbs for. This would have been nice to have as it would have allowed us to further our analysis.
- This dataset encompasses the Covid pandemic when travel was heavily impacted and therefore skews the results.

Data Ethics

As this dataset is anonymized and there is no specific location data, there are no obvious ethical concerns present.

Questions to Explore

- To what degree do weekend and weekday rates differ? Does this change depending on the city?
- Can we determine important determinants of Airbnb prices in order to predict listing prices, provide insights into areas of opportunity for improvement, or assess the effectiveness of existing policy changes concerning vacation rentals?
- Is there a way we can quantify the importance of being closer to attractions through the difference between Airbnb prices closer to attractions and those farther away?
- Are there any city differences (e.g. in policy) that can explain disparities between city data?
- Does the location of the Airbnb (e.g. its distance from metro/attractions) matter most? Meaning, are the most expensive airbnb's on average more expensive?

Hypotheses

- On average, Airbnb's are more expensive if they are closer to metros than if they are closer to attractions
- The features that affect the prices for business airbnbs differ than those for non business airbnbs

Extra Notes

- When we do a per city break down we will need to look at outliers on a per city basis as well