Advanced Analytics & Dashboard Design
6.1: Sourcing Open Data

Data source

Open and free data source for the public.
http://insideairbnb.com

Inside Airbnb is a purpose-driven project that provides data and support on Airbnb's impact on residential communities in cities where its services operate. It was founded by Murray Cox, an artist, activist and technologist, and John Morris, a designer and artist.

This site uses open source technologies to display data about Airbnb listings, including location, price and availability, as well as additional information in the form of a blog and data community forum. Thanks to this, the community has access to data and information enabling understanding, making decisions and controlling the role of renting residential houses through the Airbnb platform.

Data sets consists of 3 files:

listings csv format
neighbourhoods csv format
reviews csv format

All data under the path: 2023-10-09 Airbnb Amsterdam/02 Data/Original data

Data policies

Inside Airbnb offers a reasonable amount of free data from the last 12 months to housing and social activists and journalists. When using this data source, you should therefore quote, do not delete, collect only necessary data and handle it appropriately from an ethical point of view. For archived data, they need to be contacted in writing to the side hostels. It can result in paying for historical data or the request can receive a refusal.

Data collection

Website Inside Airbnb uses an occupancy model to estimate how often Airbnb listings are rented out and estimate the listing's revenue. The website is not affiliated with the Airbnb platform and uses only the publicly available information.

Assumptions that were made in this database and when compiling publicly available information:
1. Assuming that all reviews are accurate and complete
2. Average the length of stay of Airbnb guests is the same across all listings and cities
3. The model limits occupancy to 70%.

This information is very important because it may indicate deficiencies, biases or errors in the analysis as well as limitations resulting from this model.

The source is as reliable as the airbnb free access data.
It should be taken into account that this analysis is based on the source of the cortex influenced by many factors and it is not a trustworthy data source.

Occupancy model

It is a statistical model used to estimate the probability of a given place being occupied by a species or other entity in science.
Occupancy models are typically based on two types of data: presence-absence data and detection probability data.
Thanks to this, it is possible to understand the location, layout and factors influencing an individual's decision to choose a given area, place or, in this case, the selected property or apartment.

Data Contents

Data from 2022 provided information about the place of residence, apartment, house for rent, person owning a given property, prices, availability, reservation information and available reservations.

After opening the data for the first time, you can also assume that you have information about other years, such as 2015, 2022, 2023.

Data limitations

The possibility of human error means transcription errors or omissions of information.
We do not have historical data, which limits the scope of information to data from the last 12 months, and at the same time limits the analysis only to that time.

The data may have missing information because this is not an official portal dealing with the collection of data for the airbnb platform.

The average of data regarding length of stay may cause inconsistency because it is an imposed assumption and not a real data transfer as it really is. So, in fact, a flat can be rented for one day or even for a year, the data does not make it clear to us

It is possible to count offers on the platform that were not successful, were removed or were artificially created by interest through reviewing, increasing or lowering the offer price.

The model limits occupancy to 70%. which is a high occupancy rate and may not be accurate for all listings. For example, some listings may be rented much less frequently, while others may be rented more frequently or even all the time.

Personal interest

I chose this data set because of its geographic data and real estate interest. And of all available places, I chose Amsterdam because I have a personal interest in gaining knowledge about this place in Europe. This is knowledge that will help you understand the process of choice and how to choose your private life when choosing a place to live and a place to rest in the form of a vacation.

Data dictionary

A data dictionary has been prepared, it is located under the path: 2023-10-09 Airbnb Amsterdam/01 Project Management

Data Profile

listings.csv
Columns: 18 Rows: 8386

Data Cleaning
Columns: 15 Rows: 8386

Column name	Type of Inconsistency	Action
host_name	-	we have from every client, therefore the column with ego name is not needed and has been omitted in the new dataframe to prevent violations of private and ethical data, providing more security and anonymity
neighbourhood_group	NaN 8386	the column has no information, no values except Nan, it is not taken to the new dataframe
last_review	Column has 7556 from 8386. 830 missing values.	-
license	-	The client's private licence number is not needed for any aspect of the analysis

Neighbourhoods.csv
Columns: 2 Rows: 22

No changes in this column because it is already enabled in listings.csv, so it's not taken into account.

reviews.csv
Columns: 2 Rows: 381041

Due to the difference in the number of rows, this table is not connected to the rest. Each of the extras will be prepared and will include information about the customers' feelings.

Descriptive statistics

listings.csv

Column	Count	Mean	Min	25%	50%	75%	Max
price	8386	254.49	18	150	207	300	27857
minimum_nights	8386	5.05	1	2	3	4	1001

number_of_reviews	8386	45.44	0	3	10	36	3199
reviews_per_month	7556	1.17	0.01	0.27	0.58	1.24	120.11
calculated_host_listings_count	8386	1.83	1.0	1.0	1.0	1.0	27
availability_365	8386	82.82	0	0	18	142	365
number_of_reviews_ltm	8386	10.85	0	0	3	8	1689

Questions to explore

1. Which part of the city is most popular and which is the least popular?
2. What price range per night is appropriate to make a profit
3. What is the average rent?
4. Which real estate properties are the most popular, top 100 ?
5. How often are facilities in the top 100 rented? How many days a year are they rented?
6. In what part of the city is it best to buy real estate for sub-letting on the platform and which one?

So the goal is to understand customers' choices based on the information given to us, to find the best options, so that future investors, private individuals or people who already have a profile on Airbnb will be able to benefit from the information, as well as customers looking for the best rental options in Amsterdam.