Dataset : “Airbnb Amsterdam”

Data Source

Summary:

a. Data sourced from: https://www.kaggle.com/datasets/erikbruin/airbnb-amsterdam/data
b. Type of Data : Open Source
c. Owner : data sourced from insideairbnb.com (December 6th 2018)
d. The usability of this dataset has been rated as 10 out of 10 on Kaggle by users. The dataset is also updated frequently, on an annual basis.

● Contains at least 2 continuous variables (excluding index or ID variables, dates, years, etc.)
● Contains at least 2 categorical variables (excluding index or ID variables, dates, years, etc.)
● Contain at least 1,500 rows
● Include a geographical component with at least 2 different values

Data Profile

The final columns are given below.

Column Name	**Data Type**	**Description**
host_response_time	String	How long it takes the host to usually respond
host_response_rate	Float	Response rate of the host
host_is_superhost	String (Binary)	Is the host a superhost or not?
host_listings_count	Float	Number of listings by the host
host_has_profile_pic	String (Binary)	Indicator variable - has profile picture or not?
host_identity_verified	String (Binary)	Indicator variable - is the identity verified or not?
street	String	Street of the location
neighbourhood_cleansed	String	Neighbourhood of the location
city	String	City of the location
smart_location	String	Location

country_code	String	Country Code of the location
country	String	Country
latitude	Float	Latitude of the location
longitude	Float	Longitude of the location
is_location_exact	String	Indicator if the location is exact
property_type	String	Property Type
room_type	String	Type of room
accommodates	Integer	Number of people who can be accommodated
bathrooms	Integer	Number of bathrooms
bedrooms	Integer	Number of bedrooms
beds	Integer	Number of beds
bed_type	String	Type of bed
price	Float	Price of the apartment
guests_included	Integer	Number of guests who can be included
extra_people	Float	Cost for extra guests
minimum_nights	Integer	Minimum number of nights
maximum_nights	Integer	Maximum number of nights
has_availability	String (Binary)	Indicator for the availability of the place
availability_30	Integer	Number of days available in a 30 day period
availability_60	Integer	Number of days available in a 60 day period
availability_90	Integer	Number of days available in a 90 day period
availability_365	Integer	Number of days available in a 365 day period
number_of_reviews	Integer	Number of reviews for the place
review_scores_rating	Integer	Score (out of 100)

review_scores_accuracy	Integer	Score (out of 10)
review_scores_cleanliness	Integer	Score (out of 10)
review_scores_communication	Integer	Score (out of 10)
review_scores_location	Integer	Score (out of 10)
review_scores_value	Integer	Score (out of 10)
requires_license	String (binary)	Does the place require a licence? Indicator variable
instant_bookable	String (binary)	Is the place instantly bookable?
is_business_travel_ready	String (binary)	Is the place a business travel ready?
require_guest_profile_picture	String (binary)	Indicator if the host requires any guest profile picture
require_guest_phone_verification	String (binary)	Indicator if the host requires any guest phone verification

Data Cleaning

- The original dataset ‘listings_details’ consists of irrelevant data such as identifiers and text descriptions and URLs. Since we are not doing a sentiment analysis, such columns will be discarded.
- Columns having more than 50% of their values missing were dropped or removed
- The host response rate column has a % symbol at the end of the values. This was removed from the values and type-cast to float32.
- The Price column has a $ symbol in front of the numeric values. This was removed before type-casting to float 32. Since all the currency is recorded in dollars, there is no need for currency conversion.
- Columns having indicator values such as ‘requires_license’, ‘instant_bookable’, etc were converted to ‘True’ or ‘False’ for the purpose of Exploratory Analysis.
- The columns ‘bedrooms’, ‘bathrooms’ and ‘beds’ have 5, 3 and 3 values missing. These values were imputed with the median values of the respective columns.
- No duplicated records were found in the dataframe.
- The cleaned dataset was exported to a .csv file

Summary Statistics for numeric columns

	host_response_rate	host_listings_coun	host_total_listings_coun	latitude	longitude	accommodates	bathrooms	bedrooms	beds	price

		t	t							
mean	93.994118	7.798340	7.798340	52.366598	4.887254	2.899538	1.142234	1.415826	1.861181	156.556122
std	16.664925	32.862384	32.862384	0.014438	0.030047	1.336813	1.085288	0.878994	1.427219	120.822784
min	0.000000	0.000000	0.000000	52.289274	4.763264	1.000000	0.000000	0.000000	0.000000	0.000000
25%	100.000000	1.000000	1.000000	52.356899	4.865501	2.000000	1.000000	1.000000	1.000000	99.000000
50%	100.000000	1.000000	1.000000	52.366017	4.886337	2.000000	1.000000	1.000000	1.000000	130.000000
75%	100.000000	2.000000	2.000000	52.375435	4.904097	4.000000	1.000000	2.000000	2.000000	180.000000
max	100.000000	698.000000	698.000000	52.424641	5.010515	17.000000	100.500000	12.000000	32.000000	5040.000000

	availability_90	availability_365	number_of_reviews	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_communication	review_scores_location	review_scores_value
mean	25.853720	85.002942	32.542350	95.074296	9.699769	9.508197	9.796343	9.491803	9.169714
std	29.342725	113.595789	54.342001	6.198066	0.624701	0.778073	0.581490	0.670084	0.755774
min	0.000000	0.000000	1.000000	20.000000	2.000000	2.000000	2.000000	2.000000	2.000000
25%	0.000000	2.000000	6.000000	93.000000	10.000000	9.000000	10.000000	9.000000	9.000000
50%	12.000000	24.000000	15.000000	97.000000	10.000000	10.000000	10.000000	10.000000	9.000000
75%	51.000000	135.250000	34.000000	99.000000	10.000000	10.000000	10.000000	10.000000	10.000000
max	90.000000	365.000000	695.000000	100.0000	10.000000	10.000000	10.000000	10.000000	10.000000

Data Limitations and Ethics

The original dataset has some sensitive, personally identifiable information regarding the Host. These PII columns were removed from further analysis. The dataset was scraped from the official Airbnb website and therefore, the trustworthiness nature of the dataset can be confirmed and ethically, there is no issue with working with this data. Also, these listings have a column which would help assert the exactness of the location of the Airbnb apartment.

We do not have much information regarding the time it was collected. So, we don't know for sure if any holidays would have had any impact on any price surge. There could be Selective Bias as only selective streets/ neighbourhoods in Amsterdam would have been recorded. Exploratory Analysis is required to further confirm if this is true. Amsterdam is the dominant city recorded. There are other cities like Jordaan and Diemen which are under-represented. The same logic goes with room type where only 33 out of almost 9,500 rooms are shared. There is a room with a price of 5040 dollars in the dataset, where the mean price is only about 120 dollars. This room could be an outlier, or the dataset is limited to only budget friendly and middle-income level customers.

Potential questions that could be asked

- Do all hosts who have listings in the city of Amsterdam are prompt with responses to prospective customers?
- Does the geography of the city have anything to do with the number of listings, i.e., are listings concentrated in downtown Amsterdam or are the outskirts of the city just as promising?
- Is there any underlying linear relationship between the price of a listing and the number of bedrooms in the listing?
- Is there any segmentation possible based on the pricing of the listing? How many people can each segment accommodate?
- What are the limitations of this dataset? What can be done to estimate a more accurate prediction of pricing?
- How does the pricing structure relate to the availability of the listings for long-term stays, like a month?