# Data Source

## Airbnb Amsterdam

About Dataset

It gives a snapshot of the Airbnb Amsterdam advertisements situation on December 6th, 2018

Data Source

The data is trustworthily, provided from
http://insideairbnb.com/amsterdam/
https://www.kaggle.com/datasets/erikbruin/airbnb-amsterdam/data

Data Collection Method

The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site.

Data Content

Calendar - The calendar has 365 records for each listing. It specifies the whether the listing is available on a particular day (365 days ahead), and the price on that day.

Listings - A listing is basically an advertisement. This file holds the most useful variables that can be used visualizations.

Listings-details - This file holds the same variables as the listing file plus 80 additional variables.

Neighbourhoods - Simple file with the Dutch names of the neighborhoods

Neighbourhoods.geojson - This is the shape file that can be used in conjunction with interactive maps (such as Leaflet for R of the Python folium package).

Reviews - This is a simple file that can be used to count the number of reviews by listing (for a specific period).

Reviews_details - This file holds the full details of all reviews, and can also be used for instance for text mining.

Data Limitation

Data is limited to December 6th 2018

Data Relevancy

The data shows the geographic and daily records for each listening advertisement in Amsterdam.

Why this data set have been chosen?.

This database focuses on marketing, and it is an area in which I hope to get a job as an analyst.

According to Kaggle.com it has 100% usability.

# Data Profile

**Listings Data Set**

Reviewing variables

Shape (20030, 16)

Data columns (total 16 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 20030 non-null | int64 |
| 1 | name | 19992 non-null | object |
| 2 | host_id | 20030 non-null | int64 |
| 3 | host_name | 20026 non-null | object |
| 4 | neighbourhood_group | 0 non-null | float64 |
| 5 | neighbourhood | 20030 non-null | object |
| 6 | latitude | 20030 non-null | float64 |
| 7 | longitude | 20030 non-null | float64 |
| 8 | room_type | 20030 non-null | object |
| 9 | price | 20030 non-null | int64 |
| 10 | minimum_nights | 20030 non-null | int64 |
| 11 | number_of_reviews | 20030 non-null | int64 |
| 12 | last_review | 17624 non-null | object |
| 13 | reviews_per_month | 17624 non-null | float64 |
| 14 | calculated_host_listings_count | 20030 non-null | int64 |
| 15 | availability_365 | 20030 non-null | int64 |

Descriptive statistical analysis

| | id | host_id | price | mininights | num_rws | rws_x_month | c_host_lgs_count | a_365 |
|---|---|---|---|---|---|---|---|---|
| **count** | 20030 | 20030 | 20030 | 20030 | 20030 | 17624 | 20030 | 20030 |
| **mean** | 15417250 | 48685700 | 152 | 3 | 22 | 1 | 5 | 60 |
| **std** | 8569404 | 56496350 | 146 | 13 | 43 | 1 | 23 | 104 |
| **min** | 2818 | 3159 | 0 | 1 | 0 | 0 | 1 | 0 |
| **25%** | 8188423 | 8093516 | 96 | 2 | 3 | 0 | 1 | 0 |
| **50%** | 15630490 | 23694500 | 125 | 2 | 8 | 1 | 1 | 3 |
| **75%** | 22025770 | 68275350 | 175 | 3 | 22 | 1 | 1 | 67 |
| **max** | 30580410 | 229361200 | 8500 | 1001 | 695 | 12 | 208 | 365 |

Missing data

Neighbourhood_group has not records. Column supressed.

Name column has 38 records missing. Changed to 'unknown'

Host_name column has 4 records missing. Changed to 'unknown'

last_review 2046 missing data. Changet to 'never reviewed'

reviews_per_month 2046 missing data. Changed to '0'

**Listing_details Data Set**

Reviewing variables

Shape (20030, 96)

Data columns (total 96 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 20030 non-null | int64 |
| 1 | listing_url | 20030 non-null | object |
| 2 | scrape_id | 20030 non-null | int64 |
| 3 | last_scraped | 20030 non-null | object |
| 4 | name | 19992 non-null | object |
| 5 | summary | 19510 non-null | object |
| 6 | space | 14579 non-null | object |
| 7 | description | 19906 non-null | object |
| 8 | experiences_offered | 20030 non-null | object |
| 9 | neighborhood_overview | 13257 non-null | object |
| 10 | notes | 9031 non-null | object |
| 11 | transit | 13635 non-null | object |
| 12 | access | 12227 non-null | object |
| 13 | interaction | 11972 non-null | object |
| 14 | house_rules | 12571 non-null | object |
| 15 | thumbnail_url | 0 non-null | float64 |
| 16 | medium_url | 0 non-null | float64 |
| 17 | picture_url | 20030 non-null | object |
| 18 | xl_picture_url | 0 non-null | float64 |
| 19 | host_id | 20030 non-null | int64 |
| 20 | host_url | 20030 non-null | object |
| 21 | host_name | 20026 non-null | object |
| 22 | host_since | 20026 non-null | object |
| 23 | host_location | 19991 non-null | object |
| 24 | host_about | 11803 non-null | object |
| 25 | host_response_time | 10547 non-null | object |
| 26 | host_response_rate | 10547 non-null | object |
| 27 | host_acceptance_rate | 0 non-null | float64 |
| 28 | host_is_superhost | 20026 non-null | object |
| 29 | host_thumbnail_url | 20026 non-null | object |
| 30 | host_picture_url | 20026 non-null | object |
| 31 | host_neighbourhood | 14222 non-null | object |
| 32 | host_listings_count | 20026 non-null | float64 |
| 33 | host_total_listings_count | 20026 non-null | float64 |
| 34 | host_verifications | 20026 non-null | object |
| 35 | host_has_profile_pic | 20026 non-null | object |
| 36 | host_identity_verified | 20026 non-null | object |
| 37 | street | 20030 non-null | object |
| 38 | neighbourhood | 18377 non-null | object |
| 39 | neighbourhood_cleansed | 20030 non-null | object |

40 neighbourhood_group_cleansed 0 non-null float64
41 city 20026 non-null object
42 state 19903 non-null object
43 zipcode 19164 non-null object
44 market 19988 non-null object
45 smart_location 20030 non-null object
46 country_code 20030 non-null object
47 country 20030 non-null object
48 latitude 20030 non-null float64
49 longitude 20030 non-null float64
50 is_location_exact 20030 non-null object
51 property_type 20030 non-null object
52 room_type 20030 non-null object
53 accommodates 20030 non-null int64
54 bathrooms 20020 non-null float64
55 bedrooms 20022 non-null float64
56 beds 20023 non-null float64
57 bed_type 20030 non-null object
58 amenities 20030 non-null object
59 square_feet 406 non-null float64
60 price 20030 non-null object
61 weekly_price 2843 non-null object
62 monthly_price 1561 non-null object
63 security_deposit 13864 non-null object
64 cleaning_fee 16401 non-null object
65 guests_included 20030 non-null int64
66 extra_people 20030 non-null object
67 minimum_nights 20030 non-null int64
68 maximum_nights 20030 non-null int64
69 calendar_updated 20030 non-null object
70 has_availability 20030 non-null object
71 availability_30 20030 non-null int64
72 availability_60 20030 non-null int64
73 availability_90 20030 non-null int64
74 availability_365 20030 non-null int64
75 calendar_last_scraped 20030 non-null object
76 number_of_reviews 20030 non-null int64
77 first_review 17624 non-null object
78 last_review 17624 non-null object
79 review_scores_rating 17391 non-null float64
80 review_scores_accuracy 17381 non-null float64
81 review_scores_cleanliness 17383 non-null float64
82 review_scores_checkin 17369 non-null float64
83 review_scores_communication 17378 non-null float64
84 review_scores_location 17370 non-null float64
85 review_scores_value 17371 non-null float64
86 requires_license 20030 non-null object

87 license 9 non-null object
88 jurisdiction_names 19358 non-null object
89 instant_bookable 20030 non-null object
90 is_business_travel_ready 20030 non-null object
91 cancellation_policy 20030 non-null object
92 require_guest_profile_picture 20030 non-null object
93 require_guest_phone_verification 20030 non-null object
94 calculated_host_listings_count 20030 non-null int64
95 reviews_per_month 17624 non-null float64

Dropped Columns

'listing_url', 'scrape_id', 'last_scrape', 'host_id', 'host_url', 'host_name', 'host_since', 'host_location', 'host_about', 'host_response_time', 'host_response_rate','host_acceptance_rate','notes', 'transit', 'interaction', 'thumbnail_url', 'medium_url', 'xl_picture_url', ''picture_url", 'host_picture_url', 'first_review', 'last_review', 'jurisdiction_names', 'instant_bookable', 'is_business_travel_ready', 'require_guest_profile_picture', 'require_guest_phone_verification', 'host_thumbnail_url', 'host_neighbourhood', 'host_listings_count', 'host_total_listings_count', 'host_verifications', 'host_has_profile_pic', 'host_identity_verified', 'summary', 'listing_url', 'description', 'house_rules', 'picture_url', 'market', 'calendar_last_scraped', 'reviews_per_month', 'neighbourhood_group_cleansed', 'square_feed', 'license'.

**Calendar Data Set**

Reviewing variables

Shape (7310950, 4)

Data columns (total 4 columns):

| # | Column | Dtype |
|---|---|---|
| 0 | listing_id | int64 |
| 1 | date | object |
| 2 | available | object |
| 3 | price | object |

Changes

Date Column: Changed Dtype to datetime64[ns].

Price column: Eliminated "$" symbol and changed the Dtype to float64.

Descriptive statistical analysis

| | listing_id | date | price |
|---|---|---|---|
| **count** | 7310950 | 7310950 | 1200071 |
| **mean** | 15417250 | | 208 |
| **min** | 2818 | 6-Dec-18 | 9 |

| | | | |
|---|---|---|---|
| **25%** | 8187901 | | 109 |
| **50%** | 15630490 | | 150 |
| **75%** | 22026120 | | 240 |
| **max** | 30580410 | 6-Dec-19 | 8500 |
| **std** | 8569190 | | 281 |

Missing values

Price has 83 % null data, but it is related to the availability of the listing. (f/t)

**Reviews Data Set**

Reviewing variables

Shape (431830, 2)

Data columns (total 2 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | listing_id | 431830 non-null | int64 |
| 1 | date | 431830 non-null | object |

Changes

Date Column: Changed Dtype to datetime64[ns].

**Reviews_details Data Set**

Reviewing variables

Shape (431830, 6)

Data columns (total 6 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | listing_id | 431830 non-null | int64 |
| 1 | id | 431830 non-null | int64 |
| 2 | date | 431830 non-null | object |
| 3 | reviewer_id | 431830 non-null | int64 |
| 4 | reviewer_name | 431830 non-null | object |
| 5 | comments | 431296 non-null | object |

Changes

Date Column: Changed Dtype to datetime64[ns].

Missing data

comments column has 534 missing data. Change to 'No comments given'

**Neighbourhoods Data Set**

Reviewing variables

Shape (22, 2)

Data columns (total 2 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | neighbourhood_group | 0 non-null | float64 |
| 1 | neighbourhood | 22 non-null | object |

No data set needed or useful.

**Limitations of the Dataset**

Data Completeness: Some fields like for example 'price' may have missing or incomplete values, which can impact analysis.

Self-Reported Data: Some information like for example 'price', 'listing_name', 'room_type' is provided by hosts and may not be fully accurate.

No Demand Data: The dataset includes listing availability but does not include actual booking transactions, so it's hard to measure real occupancy rates.

Bias in Reviews: People who leave reviews tend to have extreme opinions, very positive or negative, leading to potential bias in sentiment analysis.

Limited Guest Information: The dataset does not include detailed guest demographics, which could be useful for understanding traveler behavior.

**Ethical Considerations**

Privacy Concerns: Some fields in the dataset (e.g., host names, location data) may contain personal information, raising data privacy concerns.

Neighborhood Impact: Airbnb has been criticized for contributing to housing shortages and rent increases in major cities. Analysis of this dataset should consider Airbnb's role in Amsterdam's housing market.

Data Usage Transparency: If presenting findings from this dataset, it's important to disclose its limitations and avoid making misleading conclusions.

# Key Questions

**Clarifying Questions**

How are neighborhoods defined in the dataset? Do they align with official city boundaries?

Does the dataset differentiate between different types of Airbnb listings? For example: entire home vs. private room

**Adjoining Questions**

How do Airbnb listings compare to hotel prices in Amsterdam?

Are certain neighborhoods more popular for Airbnb listings than others? If so, why?

How does seasonality affect Airbnb prices and availability?

What factors influence a listing's price the most? For example: amenities, location, reviews.

**Funneling Questions**

Do listings with more reviews tend to have higher occupancy rates?

Is there a correlation between the price of a listing and its customer rating?

Are hosts with multiple listings pricing them differently compared to single-listing hosts?

What are the common characteristics of the most expensive Airbnb listings in Amsterdam?

Do highly rated listings have certain amenities in common? For example: WiFi, kitchen, free parking.

**Elevating Questions**

What strategies could hosts use to improve their listing's visibility and profitability?

How could this analysis be expanded to compare Amsterdam's Airbnb market with other major cities?