# Airbnb Prices in European Cities

## Sourcing the Data:

**1. Reason for choosing:**

I am keenly interested in these datasets as they allow me to delve deeper into the dynamics of rental prices in major cities across Europe. Europe, one of the most sought-after travel destinations in recent years, provides a rich tapestry of data from which to draw insights. I am inquisitive to uncover any disparities in rental prices among these cities. I want to understand if factors such as proximity to metro stations, major attractions, and the property's geographical location impact these prices. Furthermore, I am eager to explore if the pricing fluctuates depending on the days of the week, specifically if there are any noticeable differences between weekends and weekdays.

**2. Data Source: Link**

The datasets are taken from Kaggle, which makes them an external data source. Besides, these datasets originally came from Zenodo.org. Zenodo is a reputable platform for sharing research outputs; the original creators of this dataset also used them to publicise an issue on ScienceDirect, which is a reputable platform that hosts a vast collection of scientific and academic research articles, journals, and papers. Overall, the datasets are trustworthy.

**3. Data Collection:**

Using web-scraping, an automated experiment was conducted to collect Airbnb offers in 10 major European cities. The offers were collected 4-6 weeks in advance for two people and two nights, including weekdays and weekends. The analysis excluded listings outside city borders and those accommodating more than six people. The authors also used TripAdvisor data to measure the attractiveness of neighbourhoods.

**4. Data Content:**

The datasets include ten big cities (Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, London, Paris, Rome, Vienna). Each town also has two separate datasets, weekends and weekdays, to examine the price at different times of the week. That makes a total of 20 files. The datasets were collected in 2021, which makes them relatively new and valuable to the analysis. All the files contain the same columns: realSum, room_type, room_shared, room_private, person_capacity, host_is_superhost, multi, biz, cleanliness_rating, guest_satisfaction_overall, bedrooms, dist, metro_dist, attr_index, attr_index_norm, rest_index, attr_index_norm, lng, lat (these columns will be explained more clearly in the profile section).

**5. Data Limitation and Ethics:**

It should be noted that the dataset might contain only some Airbnb listings in the stipulated European cities, potentially causing biases in any derived analysis. Despite the dataset's focus on European cities, it may only encompass some area of each town, potentially constraining my results' universality. The dataset may be restricted to a specific temporal period. Given the propensity of Airbnb pricing to oscillate with seasonal variations or other influential factors, a dataset limited to a particular time frame may not faithfully capture these dynamics.

Listings may not include sensitive information such as exact addresses or details about the inferior layout, but it provides the latitude and longitude of the listing; this is a piece of private information and may excluded from the dataset. Analysing Airbnb data could reveal patterns of bias or discrimination in the platform, such as disparities in pricing or preference of locations. For these reasons, the dataset is used for analysis purposes.

**6. Data relevance:**

Despite its constraints, the dataset presents a valuable opportunity for understanding the dynamics of Airbnb listing prices and the factors that influence them.

## Data Profile:

**1. Data cleaning**

| Dataset | Missing values | Missing values treatment | Duplicates |
|---|---|---|---|
| airbnb | None | N/A | None |

**2. Data Wrangling**

| Columns dropped | Columns renamed | Comment/Reason |
|---|---|---|
| Unnamed: 0 | | Dropped due to no relevance to the project |
| attr_index | | Dropped due to no relevance to the project, used column 'attr_index_nor' instead |
| rest_index | | Dropped due to no relevance to the project, used column 'rest_index_nor' instead |
| lng | | Dropped due to PII issue |
| lat | | Dropped due to PII issue |
| | realSum' : 'price' | Unclear original name |
| | multi' : 'multi_listing' | Unclear original name |
| | biz' : 'business_listing' | Unclear original name |
| | dist' : 'city_center_dist' | Unclear original name |

**3. Data profile:**

Full data profile: Link

**4. Basic statistic**

| | price | person_capacity | multi_listing | business_listing | cleanliness_rating | guest_satisfaction_overall | bedrooms |
|---|---|---|---|---|---|---|---|
| **count** | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.00000 |
| **mean** | 279.879591 | 3.161661 | 0.291353 | 0.350204 | 9.390624 | 92.628232 | 1.15876 |
| **std** | 327.948386 | 1.298545 | 0.454390 | 0.477038 | 0.954868 | 8.945531 | 0.62741 |
| **min** | 34.779339 | 2.000000 | 0.000000 | 0.000000 | 2.000000 | 20.000000 | 0.00000 |
| **25%** | 148.752174 | 2.000000 | 0.000000 | 0.000000 | 9.000000 | 90.000000 | 1.00000 |
| **50%** | 211.343089 | 3.000000 | 0.000000 | 0.000000 | 10.000000 | 95.000000 | 1.00000 |
| **75%** | 319.694287 | 4.000000 | 1.000000 | 1.000000 | 10.000000 | 99.000000 | 1.00000 |
| **max** | 18545.450285 | 6.000000 | 1.000000 | 1.000000 | 10.000000 | 100.000000 | 10.00000 |

| city_center_dist | metro_dist | attr_index_norm | rest_index_norm |
|---|---|---|---|
| 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 |
| 3.191285 | 0.681540 | 13.423792 | 22.786177 |
| 2.393803 | 0.858023 | 9.807985 | 17.804096 |
| 0.015045 | 0.002301 | 0.926301 | 0.592757 |
| 1.453142 | 0.248480 | 6.380926 | 8.751480 |
| 2.613538 | 0.413269 | 11.468305 | 17.542238 |
| 4.263077 | 0.737840 | 17.415082 | 32.964603 |
| 25.284557 | 14.273577 | 100.000000 | 100.000000 |

- Prices range from around €34.78 to €18,545.45. The median price (50th percentile) is approximately €211.34
- Person capacity ranges from 2 to 6. The mean person capacity is approximately 3.16, indicating that the typical listing accommodates around three people.
- About 29.1% of the listings are multi-listings (have multiple listings from the same host).
- Approximately 35.0% of the listings are designated as business listings.
- Cleanliness ratings range from 2 to 10. The average cleanliness rating is 9.39 out of 10, and the data is right-skewed.
- Overall ratings range from 20 to 100. The average overall guest satisfaction rating is approximately 92.63 out of 100, and the data is right-skewed.
- The number of bedrooms ranges from 0 to 10. The average number of bedrooms is approximately 1.16. The data is lightly right-skewed
- The average distance from the city centre is approximately 3.19 kilometres.
- The average distance from the nearest metro station is approximately 0.68 kilometres.

## Defining questions:

1. What is the difference between the number of listings among the ten cities?
2. What is the difference between the price of listings among the ten cities?
3. What is the difference between the number of listings on weekends/weekdays?
   Do certain types of listings (e.g., entire homes vs. private rooms) see more fluctuation in availability between weekends and weekdays?
4. What is the difference between the price of listings on weekends/weekdays?
5. What type of rooms are most / least posted? How do the prices of different room types compare within each city?

6. How does the type of host (super host or not) vary among the ten cities? Are super hosts more prevalent in certain cities
7. Do the numbers listing of the host (multi / business listing) vary among the ten cities? Do multi-listings or business listings tend to have higher or lower average prices compared to regular listings?
8. What is the relationship between distance from the city centre, the metro and price? Are there any differences in price sensitivity to distance from the city centre between different cities?
9. How do the attraction index and restaurant index affect the price? Are there any differences in price sensitivity to the availability of the restaurants between different cities?