JESSICA GATTA

Italian Airbnb Project — Data Source & Profile

Data Source

Summary of the Data Source:
I am using the **Italian Airbnb dataset** available on Kaggle (Italian Airbnb Dataset). The dataset contains detailed information on Airbnb listings across multiple Italian cities, including:

- Listing details (ID, property type, number of beds/bedrooms, bathrooms, maximum guests)
- Pricing information
- Host details (host experience, superhost status, number of listings)
- Ratings and reviews (total reviews, ratings scores, reviews per month)
- Geospatial data (neighborhood, city, coordinates)
- Date of scraping and season

The dataset contains **282,047 rows** and **26 columns**, meeting the project requirement of more than 1,500 rows and multiple continuous and categorical variables.

Data Collection

The dataset was **collected from Airbnb listings in Italy** and made publicly available on Kaggle. The data appears to have been scraped from Airbnb's public listings, including metadata such as host status, ratings, and property characteristics. No private user information (e.g., names, addresses) is included.

Data Limitations

- **No missing values** were found in the dataset.
- Certain columns, such as host_is_superhost, were stored as strings and required mapping to binary values.
- Listings may be duplicated over time, so duplicates had to be removed for accurate analysis.
- The dataset only includes publicly listed properties at the time of scraping and may not reflect the full Italian Airbnb market.
- Price and review metrics are subject to variability over time and may have seasonal or temporal biases.

JESSICA GATTA

Why This Data

I chose this dataset because I am an Airbnb host in Naples, and I wanted to analyze data that is **professionally relevant** and **practically useful**. Specifically, I want to explore:

- What factors make a host become a **superhost**
- How listing features, host experience, and ratings relate to **price and reviews**
- Patterns across different cities and neighborhoods

This dataset provides both **geospatial information** and **host/listing characteristics**, which are essential for the analyses required in this project (geospatial, regression, clustering, and visualization).

Ethical Considerations

- The dataset contains **no personally identifiable information (PII)** such as host or guest names.
- I will avoid disclosing specific addresses or private details in my analysis or dashboard.
- Any insights drawn will be generalized and used for **analytical purposes only**, with no attempt to identify individual hosts.

Superhost Requirements

To better understand my feature engineering choices, I reviewed **Airbnb's official Superhost criteria (**What's required to be a Superhost - Airbnb Help Center**)**. A host must meet all the following over the past 12 months:

1. **High response rate** – respond to **90% or more** of new messages within 24 hours.
2. **High overall rating** – maintain an **overall rating of 4.8 or higher** from guest reviews.
3. **Minimum number of stays** – have completed at least **10 stays** or **3 reservations totaling 100+ nights**.
4. **No cancellations** – zero cancellations on confirmed reservations, except for extenuating circumstances.

Relation to My Dataset:

- My dataset contains the column host_is_superhost, which directly identifies whether a host meets Airbnb's criteria.
- Other relevant features in the dataset:
    - rating_score → can be compared against the 4.8 threshold.

  - total_reviews and reviews_per_month → proxies for the minimum number of stays.

- **Limitations:** The dataset does **not include response rate or cancellation data**, so I cannot fully verify all official criteria.

Why I Created the is_top_host Binary Feature:

I created a binary column is_top_host (0 = Host, 1 = Superhost) to:

- Simplify analysis for regression and clustering.
- Allow comparison of listing features, ratings, and pricing between Superhosts and regular hosts.
- Stay consistent with Airbnb's official definition, acknowledging that some criteria (response rate, cancellations) are not available in the dataset.

Questions to Explore

1. What factors increase the likelihood of a host becoming a superhost?
2. How does host experience (years active) affect ratings and reviews?
3. Are there differences in pricing across cities and neighborhoods?
4. What listing characteristics (number of beds, property type, max guests) are associated with higher prices?
5. Can we identify clusters of listings based on price, ratings, and geolocation?
6. How do seasonality and location impact price and demand?

Data Cleaning Summary

- **Removed duplicates** based on Listings id to retain a single record per property.
- **Made a safe copy** of the DataFrame for feature engineering to avoid warnings and accidental data modification.
- **Removed price outliers** greater than 3 standard deviations from the mean to ensure meaningful analysis.
- **Log-transformed price** to normalize the distribution.
- **Feature engineering:**
  - Created is_top_host binary column (0 = Host, 1 = Superhost)
  - Calculated host_experience_years from host_since and date_of_scraping

- **Converted column types**: dates to datetime, numeric columns verified, coordinates split into latitude and longitude.
- **Cleaned column names** to lowercase with underscores for programming convenience.
- **Checked missing values:** none were present.