Data Dictionary: Synthetic Airbnb Dataset

1. city (Categorical)

  - Definition: City in Pennsylvania where the listing is located.

  - Value Logic & Constraints: Randomly selected from 10 cities with weighted probabilities (e.g., Philadelphia: 20%, York: 6%).

2. room_type (Categorical)

  - Definition: Type of room available for rent.

  - Value Logic & Constraints: Randomly chosen from ['Entire home/apt', 'Private room', 'Shared room'] with weights [0.6, 0.35, 0.05].

3. accommodates (Numeric (Discrete))

  - Definition: Number of guests the listing can accommodate.

  - Value Logic & Constraints: Determined based on number of beds and room type. Typically beds <= accommodates <= beds+2.

4. bedrooms (Numeric (Discrete))

  - Definition: Number of bedrooms in the listing.

  - Value Logic & Constraints: For Entire homes: random integer between 1 and 5. For rooms: set to 1.

5. bathrooms (Numeric (Continuous))

  - Definition: Number of bathrooms available.

  - Value Logic & Constraints: Ranges from 0.5 to 2.5+ depending on room type and number of bedrooms.

6. beds (Numeric (Discrete))

  - Definition: Total number of beds available in the listing.

  - Value Logic & Constraints: Random based on bedrooms. Shared room: 1 bed. Entire home: beds >= bedrooms.

7. minimum_nights (Numeric (Discrete))

  - Definition: Minimum number of nights required to book the listing.

  - Value Logic & Constraints: Randomly set: 2-5 nights for entire homes, 1-3 nights for rooms.

8. number_of_reviews (Numeric (Discrete))

  - Definition: Total number of reviews the listing has received.

  - Value Logic & Constraints: Drawn from ranges (0-500) with decreasing probabilities for higher counts.

9. review_score_rating (Numeric (Continuous))

  - Definition: Average review score, scaled from 0 to 5.

  - Value Logic & Constraints: Based on number_of_reviews (e.g., higher reviews -> higher typical score). Scaled from integer 0-100 to float 0.0-5.0.

10. instant_bookable (Categorical (Binary))

  - Definition: Whether the listing can be booked instantly (1 = Yes, 0 = No).

  - Value Logic & Constraints: Randomly assigned: 60% chance of being instant bookable.

11. superhost (Categorical (Binary))

  - Definition: Whether the host is a Superhost (1 = Yes, 0 = No).

  - Value Logic & Constraints: Randomly assigned: 30% chance of being a Superhost.

12. wifi (Categorical (Binary))

- Definition: Whether the listing includes WiFi (1 = Yes, 0 = No).

  - Value Logic & Constraints: Randomly assigned: 80% chance of having WiFi.

13. kitchen (Categorical (Binary))

  - Definition: Whether the listing includes a kitchen (1 = Yes, 0 = No).

  - Value Logic & Constraints: Always 1 for entire homes; randomly assigned 0 or 1 for other room types.

14. price (Numeric (Continuous))

  - Definition: Estimated nightly price of the listing in USD.

  - Value Logic & Constraints: Base price adjusted for bedrooms, amenities, review score, and city multiplier. Random noise added (-$10 to +$15).