New York City Airbnb Open Data — Background Document

Context

Airbnb, Inc. is an American company headquartered in San Francisco, California, that operates an online marketplace for lodging, primarily homestays for vacation rentals and tourism activities. Since its founding in 2008, Airbnb has grown significantly, enabling hosts and guests to interact through its platform via website and mobile applications.

This dataset focuses specifically on Airbnb activity within New York City. It was designed as part of the **Airbnb Inside** initiative, providing valuable insights into homestay listings, pricing, availability, and guest reviews in one of the most active cities for Airbnb globally.

Dataset Overview

The **New York City Airbnb Open Data** contains comprehensive details on Airbnb operations in New York, including:

- Listings: Detailed descriptions of properties, their locations, and average review scores.
- Reviews: Unique identifiers for each reviewer and associated detailed comments.
- Calendar Data: Daily prices and availability for each listing.

The dataset was expanded by **Arian Azmoudeh** to introduce additional columns and data inconsistencies for educational purposes, particularly for practicing data cleaning techniques.

Key Features

The main features (columns) included in the dataset are:

- `id`: Unique listing identifier.
- `name`: Name/title of the listing.
- `host_id`: Unique host identifier.
- `host_name`: First name of the host.
- `host_identity_verified`: Indicates whether the host's identity is verified.
- `neighbourhood_group`: Broad area classification (e.g., Manhattan, Brooklyn).
- `neighbourhood`: Specific neighborhood within the city.
- `latitude` and `longitude`: Geolocation coordinates for mapping.
- `country`: Country where the listing is located (United States).

The data includes both structured elements (such as IDs and prices) and textual descriptions (such as reviews), making it suitable for a variety of SQL tasks including querying, filtering, aggregating, and joining tables.

Suitability for the Predict Project

This dataset fulfills the project's data requirements:

- **Interesting and Rich**: Includes detailed information on thousands of listings, hosts, reviews, and pricing data.
- **Not Too Complex**: While rich, the dataset is manageable for learners and appropriate for SQL practical tasks.
- **US-Centric**: Focuses solely on listings in New York City, United States.
- **Legally Usable**: Provided under an Open Database License with content credited to the original authors. The usage for educational purposes is permitted.
- **Realistic Business Scenario**: Students can role-play as data scientists analyzing Airbnb operations, which is relevant to common real-world data science roles in hospitality and tourism industries.

Data Source and References

- **Original Data Source:**
  Inside Airbnb
- **Extended Dataset Contributor:**
  Arian Azmoudeh — LinkedIn Profile
- **License Information:**
  Database: Open Database License; Contents: © Original Authors.
  Educational use is compliant with the data's usage policy.
- **Data Dictionary:**
  Google Sheets Data Dictionary

Additional Notes

The data set has been intentionally altered to provide opportunities for students to practice data cleaning, including addressing missing values, handling outliers, and correcting inconsistencies.

No political or sensitive topics are involved, ensuring that the dataset remains appropriate for educational and business-focused use cases.

 This version fully meets professionalstandards for submission.