Data Cleaning on Airbnb open Data

1.Id columns:

Cleaning: # Remove any duplicate entries based on 'id' and drop rows with missing 'id'.

There should be no duplicates or missing IDs as each listing should be unique.

2.Name column

Cleaning: Handle missing values, strip excess whitespace, and remove non-ASCII characters.

a. # Remove leading/trailing whitespaces(use strip())
b. # Drop rows where name is missing
c. # Remove non-alphanumeric characters

name should be meaningful and not contain unnecessary spaces or special characters.

3.Host id column

Cleaning: Check for duplicates, missing values, and ensure it's a numeric column.

a. # Convert to numeric, coerce errors to NaN
b. # Drop rows where host_id is missing

Host ID should be a numeric value, and any missing or erroneous values should be handled.

4.Host name column

Cleaning: Handle missing values and clean any inconsistencies in formatting.

a. # Remove leading/trailing whitespaces
b. # Drop rows where host_name is missing
c. # Remove non-alphanumeric characters

Missing host names should be dropped, and the names should be cleaned to avoid strange characters.

5. Neighbourhood column:

Cleaning: Handle missing values and normalize the neighborhood names.

a. # Drop rows where neighbourhood is missing
b. # Standardize the names (e.g., capitalize first letters)
c. Missing neighbourhood information should be dropped, and normalization helps with consistency (e.g., "central park" -> "Central Park").

6.Latitude and longitude:

Cleaning: Check for missing values and invalid coordinates (latitude should be between -90 and 90, longitude between -180 and 180).

a. # Drop rows where coordinates are missing
b. # Latitude should be between -90 and 90
c. # Longitude should be between -180 and 180

Geographical coordinates must be valid. Any row with invalid or missing coordinates should be dropped.

7.Clean country ,country code ,instant bookable,cancellation policy(we can also remove this column but practice you clean it)

8.room_type column:

Cleaning: Ensure consistency in the room types and handle missing values.

a. # Drop rows where room_type is missing
b. # Normalize room type values (e.g., "entire home" -> "Entire Home/Apt")

Missing values should be dropped, and normalization ensures room types are consistently formatted.

9.Clean construction year column

10.Price columns

Remove rows with non-positive prices, convert to numeric, and handle missing values.

a. # Convert to numeric, coerce errors to NaN
b. # Drop rows where price is missing

Price should be a positive number, and non-numeric values should be handled (e.g., "$100" should be converted to 100).

11.Clean service fee column

12. minimum nights column:

Cleaning: Ensure the value is positive, and handle missing or invalid data.

a. # Convert to numeric
b. # Drop rows where minimum_nights is missing
c. # Ensure minimum_nights is positive

Negative or missing values for minimum nights should be removed

13. number_of_reviews columns

Cleaning: Ensure it's numeric, handle missing values, and remove negative values.

a. # Convert to numeric
b. # Drop rows where number_of_reviews is missing
c. # Ensure reviews are non-negative

Reviews should always be non-negative, and missing values should be handled.

14. last_review

Cleaning: Convert to datetime format, handle missing values

a. # Convert to datetime
b. # Drop rows where last_review is missing

The date should be in a valid format, and any rows without a valid date should be dropped.

15. availability_365 column

Cleaning: Ensure the value is within the range [0, 365] and handle missing values.

a. # Convert to numeric
b. # Drop rows where availability is missing
c. # Ensure it's within valid range

Availability should be a valid number between 0 and 365.

15. reviews_per_month columns

Cleaning: Handle missing values and convert to numeric.

a. # Convert to numeric
b. # Drop rows where reviews_per_month is missing
c. # Ensure non-negative values

16.drop remaining columns