Data Description

1) calendar.csv

Column Name	Suggested Data Type	Description
calendar_id	Text or Integer	Unique row identifier (primary key) for each calendar entry. Often not essential for analysis—mainly ensures uniqueness.
listing_id	Whole Number	Foreign key linking to listings.listing_id. Identifies which property the calendar row applies to.
date	Date	The calendar date (format: YYYY-MM-DD) for that listing. One row per listing per date.
available	Boolean	Availability flag (convert "t"/"f" or "true"/"false" to TRUE/FALSE). Indicates whether the listing is bookable on that date.
price	Decimal Number	Base price for that listing on that date (e.g., "$125.00"). Remove "$" and commas, then convert to a numeric type so you can compute averages, totals, etc.
adjusted_price	Decimal Number	Price after any discounts or length-of-stay adjustments (e.g., "$115.00"). Also strip symbols and convert to numeric.
minimum_nights	Whole Number	Minimum number of nights required to book starting on that date (e.g., 2, 3).
maximum_nights	Whole Number	Maximum number of nights allowed if a guest books starting on that date (e.g., 30, 365).

Cleaning Tips:

1. date → change to Date type.
2. available → map "t"/"f" (or "true"/"false") to Boolean (TRUE/FALSE).
3. price & adjusted_price → remove non-numeric characters ("\$", commas), then convert to Decimal.
4. Filter out rows where price is null or zero, if those represent unavailable or erroneous entries.

5. Use this table to calculate metrics like AveragePrice, AvailabilityRate, or MonthlyRevenue per listing.

---

2) Hosts.csv

Column Name	Suggested Data Type	Description
host_id	Text	Unique identifier for each host (UUID format). Links to listings.host_id.
host_name	Text	Host's display name (e.g., "Alice"). Can be null or anonymized (e.g., "JohnDoe123"). Primarily for display—avoid using it as a key in calculations.
host_since	Date	Date when the host first joined Airbnb (format: YYYY-MM-DD). Convert to Date type to compute host tenure or generate time-based analytics.
host_location	Text	Free-text location string entered by the host (e.g., "San Francisco, CA, USA," "Berlin, Germany"). May require geocoding or parsing to extract structured state/country fields.
host_about	Text	Host's self-written biography. Often multiline text; useful for sentiment or completeness metrics. Consider computing length (number of words) or flagging whether non-blank rather than storing full text.

Cleaning Tips:

1. host_since → convert from text to Date in Power Query.
2. host_location → if you want structured geography, parse out state/country or join to a geocoded lookup. Otherwise, keep as display text.
3. host_about → you can derive a "BioLength" field (count of characters or words) or a binary "HasBio" = TRUE/FALSE.

4. Link host_id to listings.csv so you can analyze performance or ratings by host tenure or location.

3) listings.csv

Column Name	Suggested Data Type	Description
listing_id	Whole Number	Unique identifier for each property/listing. Primary key for the Listings table; used to join with calendar and reviews tables.
listing_url	Text	Full URL to the Airbnb listing (e.g., https://www.airbnb.com/rooms/123456). Useful for drill-through or reference in dashboards; not used as a filter.
name	Text	Title of the listing (e.g., "Cozy Loft in Downtown"). Used for display in visuals or tooltips.
description	Text	Full text description of the property. Often long free-text. You may extract features (e.g., "number of words," "contains the word 'luxury'") if performing text analysis. Otherwise, treat as display or drop for performance.
latitude	Decimal Number	GPS latitude of the property (e.g., 40.7128). Use along with longitude to plot points on a map visual.
longitude	Decimal Number	GPS longitude of the property (e.g., −74.0060).
property_type	Text	Category of property (e.g., "Apartment," "House," "Condominium," "Guest suite"). Use as a slicer to compare prices and occupancy by property type.
room_type	Text	Subcategory of the vacation rental (e.g., "Entire home/apt," "Private room," "Shared room," or "Hotel room"). Use to analyze price/availability differences by room type.
accommodates	Whole Number	Number of guests the listing can sleep (e.g., 2, 4, 6). Useful for segmenting listings by party size; can also derive "PricePerGuest" if desired.
bathrooms_text	Text	Free-text label indicating bathroom count (e.g., "1 bath," "2.5 shared baths," "Half-bath"). Contains numeric and text. To analyze numerically, parse out the leading number (including decimals) into a new bathrooms column.

bedrooms	Whole Number	Number of bedrooms (e.g., 1, 2, 3). May contain nulls if missing; convert the column to Whole Number. Consider grouping into buckets (e.g., “1 BR,” “2–3 BR,” “4+ BR”).
beds	Whole Number	Number of beds (e.g., 1, 2, 4). Use for analyzing sleeping capacity versus number of bedrooms.
amenities	Text	Pipe- or comma-delimited list of amenities provided (e.g., ["Wifi","Kitchen","Heating","Washer"]). To analyze, split into a separate dimension or create binary indicator columns (e.g., Amenity_Wifi = TRUE/FALSE).
host_id	Whole Number	Foreign key linking to hosts.host_id. Used to join this table to host demographics.

Cleaning Tips:

1. bathrooms_text → extract the numeric portion (including decimals) into a bathrooms decimal column. Example in Power Query M:
   = try Number.FromText(Text.BeforeDelimiter([bathrooms_text], " ")) otherwise null
2. amenities → in Power Query, use Json.Document or SplitColumnByDelimiter to create a row for each amenity, then pivot to build a binary matrix (Amenity_Wifi = 1/0, etc.).
3. latitude & longitude → convert to Decimal Number. Filter out any 0,0 outliers.
4. Use property_type, room_type, accommodates, bedrooms, and beds to segment price and availability trends.

---

4) reviews.csv

Column Name	Suggested Data Type	Description
review_id	Text	Unique identifier for each review (UUID). Primary key for the Reviews table.
listing_id	Whole Number	Foreign key linking to listings.listing_id. Identifies which property each review belongs to.
date	Date	Date when the review was posted (format: YYYY-MM-DD). Convert to Date type to analyze review trends over time (e.g., monthly or yearly).

reviewer_id	Text	Unique identifier for the guest who left the review (UUID). Use to count distinct reviewers per listing or analyze reviewer behavior (repeat guests, etc.).
reviewer_name	Text	Display name of the reviewer (e.g., “John”). Mostly for display; consider whether to anonymize or drop if not required in visualizations.
comments	Text	Full text of the review. Can be quite long. For most dashboards, you can compute derived metrics (e.g., ReviewLength = LEN([comments])) or run text analytics (sentiment, common keywords).

Cleaning Tips:

1. date → convert from text to Date in Power Query.
2. Create a new ReviewCount measure in DAX:
   ReviewCount = COUNTROWS(Reviews)

   or at the listing level:

   ReviewCountByListing = CALCULATE(COUNTROWS(Reviews),Reviews[listing_id] = MAX(Listings[listing_id]))

3. For text analytics on comments, you could extract sentiment scores or keyword flags. If not used, you can drop the column or keep as a detail in a drill-through page.
4. Link listing_id to the Listings table; join listings.listing_id → reviews.listing_id to bring review data into listing-level visuals.