0.1 Question 1: Unboxing the Data

0.1.1 Question 1a

As mentioned above, we are working with just one month of data. In the full database (which we don't have access to), tables like the `data` table have billions of rows. What do you notice about the design of the database schema above that helps support the large amount of data and minimize redundancy? **Keep your response to at most two sentences.**

Hint: There is no need to examine any data here. What is a technique learned in lecture? Define that technique.

The database uses normalization, a method that divides data into numerous related tables to prevent duplication and assure consistency. This simplifies the management of billions of rows by lowering storage requirements while ensuring data integrity.

0.1.2 Question 1d

Address the two questions below:

1. Can you uniquely determine the building given the sensor data? Why? (**Hint:** given a row in the `data` table, can you determine a **uniquely** associated row in `real_estate_metadata` table? Your answer should draw insights from 1b.)
2. Could `buildings_site_mapping.building` be a valid foreign key pointing to `real_estate_metadata.building_name`? (**Hint:** think about what kinds of columns can be a foreign key.)

Please keep your response to **exactly 1 sentence for each subpart and format your answer like so:**

1. YOUR ANSWER
2. YOUR ANSWER

No, you cannot determine the building uniquely from a row in the data table since numerous buildings might have the same building name in real_estate_metadata, as seen in the JSON aggregate of Question 1b.

No, buildings_site_mapping.building is not an acceptable foreign key for real_estate_metadata.building_name because building names are not unique in the metadata, and foreign keys must reference unique values.

0.2 Question 3: Entity Resolution

0.2.1 Question 3a

There is a lot of mess in this dataset related to entity names. As a start, have a look at all of the distinct values in the `units` field of the `metadata` table, which contains the units of measurement for a particular piece of metadata (you can use the ungraded code cell below or the terminal).

If you are unfamiliar with a unit of measurement, try searching for it and its abbreviation online.

What do you notice about these values? Are there any duplicates? **Limit your response to one sentence.**

Many duplication are generated by uneven formatting, such as changes in capitalization, spacing, and nomenclature for the same units (e.g., “F”, “°F”, and “Fahrenheit”)

```
In [22]: grading_util.run_sql("""
         SELECT DISTINCT units
         FROM metadata
         ORDER BY units;
         """)
```

```
Out[22]:      units
         0        A
         1     Amps
         2   Bottom
         3       CF
         4      CFm
         ..     …
         29      uS
         30       V
         31   Volts
         32       W
         33      Wh

         [34 rows x 1 columns]
```

0.2.2 Question 3d

Moving on, have a look at the `real_estate_metadata` table—starting with the distinct values in the `location` field! What do you notice about the spelling of some of these values? (If you're unfamiliar with these locations, search them up online.) **Keep your response to at most 1 sentence.**

Certain locations contain conflicting or inaccurate spellings, such as "PARANNSS AVE" instead of "PARNASSUS AVE," and differing abbreviations, such as "AVE" vs "AV"