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 where data is separated into its relative table to avoid any sort of redundancy and make everything more efficient, Like where sensor readings are stored into separate data tables with references to meta data tables and real estate meta data tables to prevent repeated storage of sensor data.

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

1. You can't uniquely determine the building from the sensor data because building name like "Alumni Hoouse" can show up in many locations as seen in the result of 1b.

2. No because buildings_site_mapping.building maps to multipe different rows in real_estate_metadata.building_name, violating the unique requiremtn for a foreign key reference.

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 unit values are duplicated because of inconsistent formatting with variations in capitalization and spacing.

```
In [361]: grading_util.run_sql("""
          SELECT 'YOUR CODE HERE';
          """)
```

```
Out[361]:          ?column?
          0  YOUR CODE HERE
```

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.**

Many location names in location are mispelled or corrupted like "Franciscoan" instead of San Francisco is probably due to issues with data entry