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 three sentences.

Hint: There is no need to examine any data here. What is a technique learned in lecture 16? Define that technique.

ANSWER: The data was divided into different relations to minimize redundancy. This technique is called normalization, and allows for the decomposing of relations so that we aren't processing a massive table in each query.

0.1.2 Question 1d

Do you see any issues with the schema given? In particular, please address the two questions below: - 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.) - Could `buildings_site_mapping.building` be a valid foreign key pointing to `real_estate_metadata.building_name`? (**Hint:** think about the definition / constraints of a foreign key.)

Please keep your response to **at most three sentences.**

ANSWER: We cannot uniquely determine the building given the sensor data. As we can see from 1b, there are multiple buildings with the same name, so one buliding name will lead to several rows in the `real_estate_metadata` *table.* `buildings_site_mapping.building` *is not a valid foreign key pointing to* `real_estate_metadata.building_name` *because there are buildings in* `buildings_site_mapping` *that cannot refer to a row in* `real_estate_metadata` *because it simply does not exist.*

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

ANSWER: There are a few units that could be representing the same thing but are only a letter or two off from each other or represented in different ways (like pounds and lbs), so they are recognized as distinct units.

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 these values? Keep your response to at most two sentences.

Some of the values are spelled correctly but others have mispellings so even a letter off will cause it to represent a new row. For example, there is 'FRANCISC O' and 'FRANCISC SOAN', which we can infer both represent San Francisco.