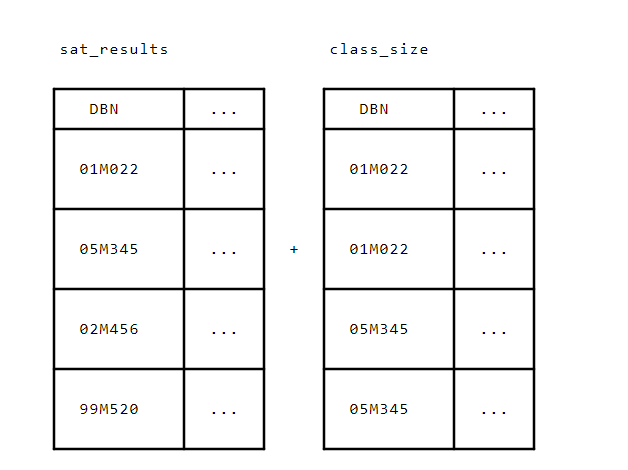
In the last mission, we began investigating possible relationships between SAT scores and demographic factors. In order to do this, we acquired several data sets about [New York City public schools](https://data.cityofnewyork.us/data?cat=education). We manipulated these data sets, and found that we could combine them all using the DBN column. All of the data sets are currently stored as keys in the data dictionary. Each individual data set is a pandas dataframe.

In this mission, we'll clean the data a bit more, then combine it. Finally, we'll compute correlations and perform some analysis.

The first thing we'll need to do in preparation for the merge is condense some of the data sets. In the last mission, we noticed that the values in the DBNcolumn were unique in the sat\_results data set. Other data sets like class\_size had duplicate DBN values, however.

We'll need to condense these data sets so that each value in the DBN column is unique. If not, we'll run into issues when it comes time to combine the data sets.

While the main data set we want to analyze, sat\_results, has unique DBN values for every high school in New York City, other data sets aren't as clean. A single row in the sat\_results data set may match multiple rows in the class\_size data set, for example. This situation will create problems, because we don't know which of the multiple entries in the class\_size data set we should combine with the single matching entry in sat\_results. Here's a diagram that illustrates the problem:

In the diagram above, we can't just combine the rows from both data sets because there are several cases where multiple rows in class\_size match a single row in sat\_results.

To resolve this issue, we'll condense the class\_size, graduation, and demographics data sets so that each DBN is unique.