In the last step, we saw a group of dataframes that looked like this:



CSD BOROUGH SCHOOL CODE                SCHOOL NAME GRADE  PROGRAM TYPE  \

0    1       M        M015  P.S. 015 Roberto Clemente     0K       GEN ED

1    1       M        M015  P.S. 015 Roberto Clemente     0K          CTT

2    1       M        M015  P.S. 015 Roberto Clemente     01       GEN ED

3    1       M        M015  P.S. 015 Roberto Clemente     01          CTT

4    1       M        M015  P.S. 015 Roberto Clemente     02       GEN ED

We can make some observations based on the first few rows of each one.

* Each data set appears to either have a DBN column, or the information we need to create one. That means we can use a DBN column to combine the data sets. First we'll pinpoint matching rows from different data sets by looking for identical DBNs, then group all of their columns together in a single data set.
* Some fields look interesting for mapping -- particularly Location 1, which contains coordinates inside a larger string.
* Some of the data sets appear to contain multiple rows for each school (because the rows have duplicate DBN values). That means we’ll have to do some preprocessing to ensure that each DBN is unique within each data set. If we don't do this, we'll run into problems when we combine the data sets, because we might be merging two rows in one data set with one row in another data set.

Before we proceed with the merge, we should make sure we have all of the data we want to unify. We mentioned the survey data earlier (survey\_all.txtand survey\_d75.txt), but we didn't read those files in because they're in a slightly more complex format.

Each survey text file looks like this:



dbn bn  schoolname  d75 studentssurveyed    highschool  schooltype  rr\_s

"01M015"    "M015"  "P.S. 015 Roberto Clemente" 0   "No"    0   "Elementary School"     88

The files are *tab delimited* and encoded with Windows-1252 encoding. An encoding defines how a computer stores the contents of a file in binary. The most common encodings are UTF-8 and ASCII. Windows-1252 is rarely used, and can cause errors if we read such a file in without specifying the encoding. If you'd like to read more about encodings, [here's](http://kunststube.net/encoding/) a good primer.

We'll need to specify the encoding and delimiter to the pandas [pandas.read\_csv()](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html" \t "_blank) function to ensure it reads the surveys in properly.

After we read in the survey data, we'll want to combine it into a single dataframe. We can do this by calling the [pandas.concat()](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.concat.html" \t "_blank) function:



z = pd.concat([x,y], axis=0)

The code above will combine dataframes x and y by essentially appending y to the end of x. The combined dataframe z will have the number of rows in x plus the number of rows in y.