# Capstone 1 Data Wrangling/Data Cleaning Steps

## Step one: Synthesize data (download data)

The activity to download, clean data types, and save the data happens in 500city\_synthesize.ipynb.

* First, I download the data from CDC socrata app to make the pulling of data reproducible on any machine or by any user.
* Next, I set the data types of the columns that needed to be updated. Percentage values and population counts were stored as a string and I changed them to numeric.
* I identified three types of information.
  + Location information: this is dimensional data that holds city, state, fips and other data. This data is associated with one uniqueid (census tract/city code/ etc).
  + Data Definitions: this has information describing what different out comes are. Examples include the question asked on the survey, shortened question text, and others. Identified by categoryid.
  + Data Columns: this stores the meat of the data in which most analysis can be done. We can join the other two tables by uniqueid/cateogryid whenever needed.
* Finally, I saved three different CSVs with those three dataframes. This enables the file sizes to be smaller, upload into a future notebook quicker, and keeps over all data file sizes smaller than storing the original document in one csv.

## Step two: Explore and clean data