The New York csv originally came with 399918 rows with 18 columns. The values in the zip code column containing a decimal place and the address being separated between building number and street. ‘DBA’ represents the name of the establishment, ‘CAMIS’ being the code assigned to the establishment and ‘RECORD DATE’ being the same throughout the data frame.

I first started by making a copy of the original data frame so that I do not manipulate it in any way. Next, I dropped any duplicates and filled in any missing value under the ***zipcode*** column with 0 and defined all the values within that column as integers. I next simplified the ***building*** and ***street*** columns into a new ***address*** column by combing the 2 columns. After organizing the columns, I did a .value\_counts() for the ‘RECORD DATE’ column and got only 08/28/2017. I unselected the RECORD DATE column because it only describes the date the inspection results recorded into the file which is not needed.

The San Francisco csv had 53973 rows with 23 columns. The latitude, longitude, location columns were filled with null values and the inspection\_id was a combination of the building id and the inspection date