Facilities De-Duplication

Goal: To identify possible duplicated facility names, and suggest potential improvements to the process that might help eliminate these.

Result: Suggestions for improvement are below.

To identify possible duplicates, I joined the facilities to itself using the facdomain, facgroup, facsubgrp, factype, address, and geometry fields. I also joined on BBL, knowing that two facility records with similar names had some likelihood of being for the same facility. I also looked for records whose ID numbers were within one of each other, as this seemed to produce a more focused result set.

Even with all of these conditions, the result set contains records that are not obviously duplicates. Nevertheless, some of them are clearly duplicates, and may be reduced by some of the following actions:

1. Standardize abbreviations. There are numerous abbreviations embedded in these strings (e.g. INC, LLC, etc), and standardization of these might help to eliminate some of the duplicates. A list of abbreviations I observed is included in this folder.
2. Additionally, format them in the same way. Decide whether they should be followed by a period or not, or preceded by a comma and a space.
3. Remove portions of the string that have more than one consecutive space.
4. Use established sources to normalize names of facilities, e.g. LPC data for historic landmarks.
5. If the name contains incorrect punctuation, correct it. There should always be a space after a comma. A facility name should not end in a comma.
6. Some facility names are followed by numbers, e.g. 'ANCHOR HOUSE, INC. SRR 1'. I'm not sure of the purpose of these numbers, but elimination of these might reduce duplicates.
7. I used SQL to identify the duplicates and it is a blunt instrument. Many of the "duplicates" in the attached list are probably not actually duplicates – most of the observation I made were through reviewing the output manually. More successful matching might be accomplished with Python fuzzy matching libraries, which I gather can determine the ratio of the strings that match.

Also attached is a list of agencies more likely to send these facility names.

Supporting Documents

Jupyter Notebook

Query Result Spreadsheet

Wednesday, April 8, 2020