Affordable Housing Dataset in Chicago

Found: https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi/data

Question: “Is there a correlation/classifier that can explain where/how many affordable housing units will be more likely?”

Is it possible to append education, and Poverty data based on Zip Code?

A. Format
    a. What is the format of this dataset, and is it usable?
        i. Csv, yes
B. Examples
    a. What type of data do these examples/observations represent?
        i. Community Area
    b. Do these data have the potential to be converted into examples/observations that can be used by a set of DM processes in RM?
        i. Yes, each represent a neighborhood
C. Attributes
    a. How many truly semantically distinct attributes are in this dataset?
    b. Try to state what type of problem(s) this dataset could address, given the attributes included in the dataset Be as specific as you can, referencing the specific attributes that make this possible.
        i. Property Type and Zip Code are the attributes most useful
        ii. Number of affordable housing units in a community area can be calculated from this dataset
            1. This is why I would be interested in expanding the dataset into education and socioeconomic factors, including mortality
            2. Education
                a. CPS performance Level
                b. Healthy Schools
                c. Safety Score
                d. Family Involvement
                e. Instruction
                f. Student attendance
            3. Census
                a. Housing Crowded
                b. Household below poverty
                c. Per capita income
D. Dataset Size
    a. Look at the examples: How many are there? Will you have to sample
        i. 77
        ii. I don’t think we will have to sample
    b. Do you have enough examples for both a training and a test set.
        i. I believe we can make a 70% 15% 15% division for training dev and testing
E. Data Prep

a. Do you have a lot of missing values?
   i. Yes. in latitude longitude, and a few in property type
   ii. Half of the education data is NDA but there are attributes that are interesting and have enough data

b. Other issues that require cleaning?
   i. I think we need to merge with education at least to have enough data to analyze, we can link by community number can link to census data

c. Normalization required?
   i. Census
      1. Data is percentages, but per capita income should be normalized
      2. CPS mostly percentages

d. Changes of data types?
   i. NDA to none

e. What kinds of preprocessing do you think might be appropriate on this dataset?
   i. Connect the data to each other
   ii. Separate class of school