Deliverable 2

This week we revamped our code that cleaned up the Owner Name and Owner Address columns. By utilizing pandas dataframes, we were able to resolve our issue with commas appearing in previous columns and produce a cleaned output file. We were also able to open the data files more efficiently, as we were having trouble working with the entirety of the dataset due to its sheer size. Then, we decided to filter out parcels that did not correspond to state owned, vacant land. Land Use Codes (luc_adj1 and luc_adj2) 91, 92, and 97 allow us to find vacant or unused government land parcels. With our updated dataset of relevant properties, we can then implement FuzzyWuzzy string matching on the owner names in order to standardize each address. We encountered some cases where the variations in state agency names made it impossible to standardize every address, and we decided to attempt to manually adjust these cases. An example of this would be owner names being "MBTA", "Mass. Bay Trans", or "Massachusetts Bay Transit Authority". Since we know that we will have problems with these cases, we can go in manually and standardize this ourselves, which will allow us to remove duplicates.

After completing the standardization, we will finally have a dataset of relevant land parcels for us to analyze values. We will utilize the Zillow API and the given land values from the dataset to determine the value of these properties by total sq. value. We will examine immediate neighbors of said properties, as well as the value of what the property consists of (a vacant/empty lot vs. an unused building). We had problems accessing the Zillow API this past weekend (servers were unavailable) but we plan on beginning preliminary analysis of property values neighboring state surplus land from our updated dataset. From there we will have a better understanding of how to gauge and ultimately estimate property values.

Questions

1. Should we be merging our dataset with the other groups, or keep it separate?
    a. If we are, how do we resolve potential merging conflicts?
    b. If not, how do we make sure there isn't any overlap between our final analyses

Immediate Next Steps

1. Filter the lands by luc codes 91, 92, 97
2. Run the standardization algorithm on all sections of the dataset
3. Combine the datasets with the standardized names and open with pandas.
4. Return these lands by decreasing total sq foot value
    a. Utilize the Zillow API or land valuations given in the dataset