Team The Conquistadors
Kaijie Zhou, Janice He, Tommy Lam, Athina Said, Murtaza Moiyadi, Manuja DeSilva

State Surplus Project Deliverable 1

This week we focus on reading the data and cleaning the data for the Massachusetts Land Parcel Database. We mostly focus on just cleaning up the OwnerName and OwnerAddress columns. Due to the size of our data, we first split the data into smaller segments. We then sort the Owner Address columns using an function we found online that help to sort address base on street names, the function also group street address with prefixes together. (ex: South, North, East, West). After we group the data by addresses, we perform a standardization on the owner names by using FuzzyWuzzy for string matching and dictionary to records the scores of matching. The owner name with the highest score from every unique address is use as the standardize name for that address.

However there are still more data cleaning than we have to do. Right now the output of our cleaning algorithm is just the 2 columns from the dataset, which means we need to add in other columns into account as well for our next step. Also we encounter a problem with having commas in other columns before the owner name and owner address columns, which lead to misformat of the owner address and owner names. We need to further fine tune our cleanup file to achieve a better result. However our current result is able to clean up approximately 75% of the data, and 1 sample output file for the first segment of our data is available at kaijiez/StateSurplusLand branch “mergeOwnerNames” on github.