I. Data Cleaning

1. Remove Exclusions
   - Remove all the owners that appears to be government. It is not necessary to look into government-owned properties.
   - Look at the most common owners and modify the remove list

2. Fill in missing zip codes
   - Store all the missing (NaN) values in the 'missing_zips' variables

   Assuming that the data is sorted by zip, we conduct these two steps:
   - If the neighbor zip values (before and after) are the same, fill in the missing zip with that value
   - For any remaining missing values, fill in the Zip value from the previous row

3. Fill in missing AVTOT, AVLAND, FULLVAL
   - Replace any null values of these 3 fields with 0
   - Create temporary dataframe for values that are not 0 and calculate means for AVATOT, AVALAND, FULLVAL for each unique value in TAXCLASS
   - Replace 0 values with the corresponding mean value calculated

4. Fill in missing STORIES
   - Store all the missing (NaN) values in a temporary dataframe
   - Replace the null values in STORIES with 0
   - Calculate the mean of STORIES for each unique value in TAXCLASS
   - Replace 0 values with the corresponding mean value calculated

5. Fill in missing LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT
   - Because these fields do not have null values, we replace any value '1' or '0' with NaN
   - Calculate the mean of each of these fields for each unique value in TAXCLASS
   - Replace any null value with the corresponding mean value calculated

In summary, imputation for this dataset has been done in 2 ways:

- For Zip cope, we fill in by neighboring values
- For the other fields, we fill in by averaging the values of each column by TAXCLASS and replace the missing value with the average value. Each field is done differently since missing values are under different forms for each field (NaN or 0 or 1)

II. Variable Creation

a. Variable List

Description of Variables	# Variables Created
Lot size (ltsize) Calculate the total size of the lot by multiplying LTFRONT and LTDEPTH	1
Building size (bldsize) Calculate the total size of the lot by multiplying BLDFRONT and BLDDEPTH	1
Building volume (bldvol) Calculate the total volume of the building by multiplying bldsize and STORIES	1
R1, r2, r3 Calculate the ratio between market value (FULLVAL) vs the 3 variables created above (ltsize, bldsize, bldvol)	3
R4, r5, r6 Calculate the ratio between the land value (AVLAND) vs the 3 variables created above (ltsize, bldsize, bldvol)	3
R7,r8,r9 Calculate the ratio between the total value (AVTOT) vs the 3 variables created above (ltsize, bldsize, bldvol)	3
Inverse (r1inv, r2inv, … r9inv) These are the inverse of the corresponding variables r1, r2, ..., r9, respectively, with a small epsilon added to avoid division by zero.	9
_zip5 Create the grouped averages of these 18 variables, grouped by ZIP. Divide each of the 18 ratio variables by the two scale factors from these groupings	18
_taxclass Create the grouped averages of these 18 variables, grouped by TAXCLASS. Divide each of the 18 ratio variables by the two scale factors from these groupings	18
Value_ratio Compare the 3 value measures (FULLVAL/(AVLAND + AVTOT) with a conditional transformation	1
**** CREATED NEW** **ZScore (**FULLVAL, AVLAND, AVTOT) Calculate the z-score of the value measures variable (FULLVAL, AVLAND, AVTOT) within each TAXCLASS group.	3
Total Number of Variables Created	**61**
TOTAL VARIABLES FOR MODELING **Dropping ltsize, bldsize, bldvol after variables created*	**58**

b. Logic why these variables are useful in measuring unusualness:

- **r1 to r9 and r1inv to r9inv:**

Provide insights how the values of the property compare to the size and volume. Unusual values of these ratios could help detect any overevaluation or undervaluation, which indicates potential fraud.

- **_zip5 and _taxclass:**
  Unusual values of these scale factors indicate that the ratio is significantly different from the expected range, which demonstrates inconsistencies in the valuation for that geographic (ZIP) or tax class (TAXCLASS(group)

- **value_ratio:**
  This ratio can spot out properties with unusually low or high assessed value vs land and total values, which could mean they are undervalued or overvalued for tax purposes.

- **Zscore:**
  This measures how many standard deviations a data is away from the mean. High z-scores indicate unusually low or high values compared to the belonging tax class, which indicates potential anomalies in land and values.

III. Notebook (attached)