Homework 7 - Data Cleaning and Imputation

1. The logic and results of the exclusions

The data cleaning part primarily focuses on excluding certain properties from the analysis which are not deemed useful for the fraud detection model.

Here's a breakdown of the logic and results of the exclusions from the code:

A. Removing records with easement type as government (U):

An easement refers to the right to use someone else's land for a specific purpose. For example, utilities often have easements on properties to run power lines. Records with easement type U are removed from the data, as these properties are typically government-owned or related and may not accurately reflect market property values. Only one record is removed in this step.

B. Removing records related to government and cemetery:

The code then removes any properties that have a name associated with government departments or cemeteries. The logic behind this is that such properties, owned by government agencies or designated as cemeteries, usually have a distinct valuation pattern that might skew the anomaly detection for general properties. A list of keywords associated with government entities and cemeteries is defined (like 'DEPT', 'DEPARTMENT', 'UNITED STATES', 'GOVERNMENT', 'GOVT', 'CEMETERY'), and the `OWNER` names are scanned for these keywords. All such properties are removed from the analysis.

It also ensures that the term 'STORES' is not included in an owner's name, presumably to avoid excluding commercial properties that might include terms like 'DEPT' in their name but are not government-owned.

The code first identifies a list of unique property owners and then goes through this list to find names that match the government or cemetery-related keywords. These matched names are then added to a `remove_list`.

By performing these exclusions, the resulting dataset is expected to consist primarily of privately owned properties, which are the main focus for fraud detection in this case. We removed 26501 records in total.

Akansha Lalwani (A59019733)

2. The logic and results for every field imputation

Field	Method of Imputation	Original Number of Missing Values	Filled Missing Values
ZIP	1. Mapping by 'staddr_boro' 2. Filling with before and after ZIPs 3. Filling with previous ZIP	20431	1. 2832 2. 9491 3. 8108
FULLVAL	1. Group by ['TAXCLASS', 'BORO', 'BLDGCL'] and fillna with mean 2. Group by ['TAXCLASS', 'BORO'] and fillna with mean 3. Group by ['TAXCLASS'] and fillna with mean	10025	1. 2718 2. 6647 3. 386
AVLAND	1. Group by ['TAXCLASS', 'BORO', 'BLDGCL'] and fillna with mean 2. Group by ['TAXCLASS', 'BORO'] and fillna with mean 3. Group by ['TAXCLASS'] and fillna with mean	10027	1. 2720 2. 6641 3. 386
AVTOT	1. Group by ['TAXCLASS', 'BORO', 'BLDGCL'] and fillna with mean 2. Group by ['TAXCLASS', 'BORO'] and fillna with mean 3. Group by ['TAXCLASS'] and fillna with mean	10025	1. 2718 2. 6647 3. 386
STORIES	1. Fill with modes grouped by ['BORO', 'BLDGCL'] 2. Group by ['TAXCLASS'] and fillna with mean	42030	1. 4108 2. 37922
LTFRONT	Groupby ['TAXCLASS', 'BORO'], fill with mean. If any nulls remain, group by ['TAXCLASS'], fill with mean.	160565	160563
LTDEPTH	Groupby ['TAXCLASS', 'BORO'], fill with mean. If any nulls remain, group by ['TAXCLASS'], fill with mean.	161656	161654
BLDFRONT	Groupby ['TAXCLASS', 'BORO', 'BLDGCL'], fill with mean. If any nulls remain, group by ['TAXCLASS', 'BORO'], then ['TAXCLASS'], fill with mean.	0 (only zero values treated as missing)	All zero values replaced
BLDDEPTH	Groupby ['TAXCLASS', 'BORO', 'BLDGCL'], fill with mean. If any nulls remain, group by ['TAXCLASS', 'BORO'], then ['TAXCLASS'], fill with mean.	0 (only zero values treated as missing)	All zero values replaced

3. The logic for all variables

Here's a brief explanation of the logic behind the creation of these variables:

1. Ratio Variables - 9

- These are variables calculated as the ratio of 3 $ value fields to each of the 3 size variables calculated above.
- They will help to find properties with unusually large values, as they would have larger ratios.
- The nine variables are: `r1`, `r2`, `r3`, `r4`, `r5`, `r6`, `r7`, `r8`, and `r9`.

2. Inverse Ratio Variables - 9

- These are the inverse (1 over) of the 9 ratio variables.
- They will help to find properties with unusually small values, as they would have larger inverse ratios.
- The nine inverse ratio variables are: `r1inv`, `r2inv`, `r3inv`, `r4inv`, `r5inv`, `r6inv`, `r7inv`, `r8inv`, `r9inv`.

3. Standardized Ratio Variables - 36

- These are variables that compare each property's primary variables `r1` through `r9` and their inverses to their geographical (grouped by ZIP) or logical (grouped by TAXCLASS) neighbors.
- The logic here is to standardize the ratio variables by group, providing more context in identifying anomalies. Each of these 18 ratio variables is divided by the respective average ratio of the group the property belongs to.
- The suffixes `_zip5` and `_taxclass` are used to indicate which group's mean was used for standardization.

4. Value Ratio - 1

- This variable indicates the appropriateness of how the three value fields relate. It is calculated as `FULLVAL` divided by the sum of `AVLAND` and `AVTOT`.
- To make it useful for detecting both unusually large and small values, it's then normalized by the mean, and the inverse is taken for values less than 1.

5. Size Ratio - 1

- This is the ratio of the building size to the lot size (`bldsize` / `ltsize`), which could indicate anomalies if the building size is disproportionately large or small compared to the lot size.

We created 56 variables in total.

4. A list of all variables with their meaning

Variable Group	Variable Names	Algebraic Form	Description
Ratio Variables	r1	FULLVAL / ltsize	Ratios of three property value variables to three size variables
	r2	FULLVAL / bldsize	
	r3	FULLVAL / bldvol	
	r4	AVLAND / ltsize	
	r5	AVLAND / bldsize	
	r6	AVLAND / bldvol	
	r7	AVTOT / ltsize	
	r8	AVTOT / bldsize	
	r9	AVTOT / bldvol	
Inverse Ratio Variables	r1inv to r9inv	1/(ri + epsilon) where i is in [1, 9]	Inverse of the nine ratio variables (r1 - r9)
Standardized Ratio Variables (grouped by ZIP)	r1_zip5 to r9_zip5, r1inv_zip5 to r9inv_zip5	ri / mean(ri) for each ZIP where i is in [1, 9] and also in [r1inv, r9inv]	Nine ratio variables and their inverses standardized by ZIP group
Standardized Ratio Variables (grouped by TAXCLASS)	r1_taxclass to r9_taxclass, r1inv_taxclass to r9inv_taxclass	ri / mean(ri) for each TAXCLASS where i is in [1, 9] and also in [r1inv, r9inv]	Nine ratio variables and their inverses standardized by TAXCLASS group
Additional Variables	value_ratio	max(FULLVAL / (AVLAND + AVTOT) / mean(FULLVAL / (AVLAND + AVTOT)), 1 / (FULLVAL / (AVLAND + AVTOT) / mean(FULLVAL / (AVLAND + AVTOT)))	Indicator of how well the three value fields (FULLVAL, AVLAND, AVTOT) relate to each other
	size_ratio	bldsize / ltsize	Ratio of building size to lot size