

Data Cleansing

Implemented in SEED V 1.2.2

10/20/2015

In-range Checking

SEED checks that values are within the minimum and maximum range specified in Table 1. For the date values, the maximum value would be the current date or the current year when the function is called.

There are two types of thresholds for inrange checking. “Error” flags invalid values for the field, such as negative values for the floor areas and EUI (Energy Use Intensity). “Warning” flags values that might indicate underperforming buildings. This threshold specifically checks for low EUI (Energy Use Intensity) and Energy Star Score.

Table 1. Thresholds for database fields

Database Field	Minimum	Maximum	Type	Units
year_built	1700	present year	Error	
year_ending	1900	present year	Error	
conditioned_floor_area	0	7 million	Error	square feet
conditioned_floor_area	100		Warning	square feet
energy_score	0	100	Error	
energy_score	10		Warning	
generation_date	1/1/1900	present date	Error	
gross_floor_area	100	7 million	Error	square feet
occupied_floor_area	100	7 million	Error	square feet
recent_sale_date	1/1/1900	present date	Error	
release_date	1/1/1900	present date	Error	
site_eui	0	1,000	Error	kBtu/sq. ft./year
site_eui	10		Warning	kBtu/sq. ft./year
site_eui_weather_normalized	0	1,000	Error	kBtu/sq. ft./year
source_eui	0	1,000	Error	kBtu/sq. ft./year
source_eui	10		Warning	kBtu/sq. ft./year
source_eui_weather_normalized	10	1,000	Error	kBtu/sq. ft./year

Data type check

SEED checks fields with the correct type of the data. For instance the “Gross Floor Area” field might have non-numeric values such as “Not Available” or “Unknown” mixed in with the valid numeric values. In another case, the “City” field might have numeric values where the program expects a non-digit string. The table below shows all of the fields available in the database and their data type. Another check is for valid date formats.

Table 2. Data type categories for each field

Numeric Data Field	String Data Field	Date Format
block_number building_count conditioned_floor_area energy_score gross_floor_area lot_number occupied_floor_area postal_code site_eui site_eui_weather_normalized source_eui source_eui_weather_normalized	address_line_1 address_line_2 building_certification city custom_id_1 district energy_alerts owner owner_address owner_city_state owner_email owner_postal_code owner_telephone pm_property_id property_name property_notes space_alerts state_province tax_lot_id use_description year_built	generation_date recent_sale_date release_date year_ending