## Data Cleaning & Preparation Summary

This project combined two datasets:
1. Student Housing Survey Data (2016–2024) — student-reported addresses and demographics
2. Building Violation Data — property violation records from a city inspection database

The goal was to analyze housing conditions for off-campus students by matching addresses and comparing student density with violation counts.

### Step 1: Load and Rename Columns

The student dataset had confusing column names like '6a. street #' and '6b. street name'. These were renamed for clarity:
- street_number
- street_name
- street_suffix
- unit_number
- zip_code

### Step 2: Clean ZIP Codes

ZIP codes in the student data were sometimes only 4 digits (e.g., '2134' instead of '02134'). We padded all ZIP codes to 5 digits using zfill(5) to ensure consistency with the violation data.

### Step 3: Standardize Address Components

We cleaned key address fields in both datasets to ensure consistent formatting:
- Stripped whitespace
- Converted to uppercase
- Removed letters from street numbers (e.g., '116H' → '116')
- Replaced missing/invalid values with blanks

### Step 4: Construct simple_address_key

We created a simplified join key using street_number, street_name, and zip_code, concatenated with spaces. This key was built for both datasets to enable merging.

### Step 5: Group Data by Address

We grouped both datasets by simple_address_key to count:
- student_count: how many students reported each address
- violation_count: how many violations were tied to each address

### Step 6: Merge Datasets

We performed a left join on simple_address_key to keep all addresses from the student data and merge in violation counts. Missing violation counts were set to 0.

### Step 7: Filter for Relevant Addresses

To focus on problematic housing, we filtered the merged dataset to only include addresses with at least one violation (violation_count > 0).