Bridging Knowledge Gaps with AI: What This Code Does and Why It Matters

---

Purpose

This Python script was developed using an AI-assisted approach (via ChatGPT, Claude, and Cursor) to automate and accelerate the cleaning of inconsistent address data – particularly for quarterly graduation file imports. It replaces a slow, manual process with a structured, semi-automated workflow that is significantly faster, more consistent, and easier to audit.

What the Code Does

1. Standardises Address Formats

- Expands common abbreviations (e.g. *Rd → Road*, *Tce → Terrace*)
- Preserves valid unit numbers (e.g. *4/26 Main Road*)
- Applies consistent formatting for `PO Box` and `Mc` prefixes
- Cleans up trailing slashes, misplaced commas, and inconsistent spacing

2. Extracts Postal Codes and Country Values

- Detects a wide range of international postal code formats
- Extracts and relocates postal codes and countries from free-text fields
- Detects New Zealand cities incorrectly placed in the Country column and reassigns them appropriately

3. Handles New Zealand–Specific Structures

- Moves misplaced suburb or city names into the correct fields
- Uses a suburb lookup list with fuzzy matching to resolve inconsistencies
- Standardises Rural Delivery (RD) codes and moves them to the Suburb field if needed

4. Preserves Original Data

- Original values are saved in an `Original Data` column to maintain traceability
- While transformation logs are not yet implemented, this structure supports future edit tracking

5. Formats Output for Post-Processing

- Final output includes structured columns:
  `Address Line 1`, `Address Line 2`, `Suburb`, `City`, `Region/Province/State`, `Postal Code`, `Country`, `Original Data`, and `Has Edits`
- Applies proper casing and formatting for improved clarity and consistency

Design Philosophy

This tool is built to address the **common, high-volume data hygiene issues** found in institutional datasets – not to handle every outlier.

Rather than attempting to code for every exception, the approach is pragmatic:
automate what can be automated, and allow a human to manage the rest.

A final visual review is still required – but the volume of manual fixes has been significantly reduced.

Impact

As demonstrated in the before–and–after output:

- Fragmented or misaligned address data is now consistently structured
- Common formatting issues are resolved automatically
- Original values are preserved via the `Original Data` field
- Manual intervention is still required, but only for a small subset of cases

While the output is not yet fully CRM-ready, it is a significant improvement – reducing a three-week manual task to just a few days.
