Case study 1 Documentation

Summarize step:

I began by importing the necessary Python libraries: pandas for data manipulation, NumPy for numerical operations, and unidecode for handling text encoding. I then read data from a CSV file and previewed the first few rows to get an overview of the dataset.

Next, I checked the data for missing values across all columns. After discovering a row with a missing address, I decided to drop it from the analysis.

To handle inconsistencies in the address text and diacritical marks, I created a function to remove these marks and convert the addresses to uppercase for consistency.

My primary focus was on addresses relevant to Da Nang (DN) city. I created a list of 10 locations – eight suburbs belonging to DN and two names representing DN itself. Using this list, I filtered the addresses, creating a new column in a new Dataframe containing 3,188 addresses belonging to DN.

For the remaining addresses not associated with DN, my goal was to identify the corresponding provinces. Following a similar approach, I listed all provinces in Vietnam except DN, removed diacritical marks, and converted the text to uppercase. Applying the same filtering process as before, I obtained a new DataFrame with the correct matching provinces for these non-DN addresses (362 rows).

Finally, I combine two sub data frames into a complete one. As a result, I obtained a new Dataframe with 3550 accurate addresses.

2. Result

Out of a total of 4479 entries, I managed to convert 3550 rows successfully. However, some entries couldn't be converted due to incorrect information provided by users. These errors mainly stem from misunderstandings/subjective reasons, such as entering suburbs instead of provinces or inputting information in the wrong column (e.g., names instead of addresses).

3. Next steps/Questions/Difficulties

To resolve this issue, we could review the unsuccessful conversions to identify common patterns or reasons for failure and consider implementing user-friendly input validation to catch common errors.