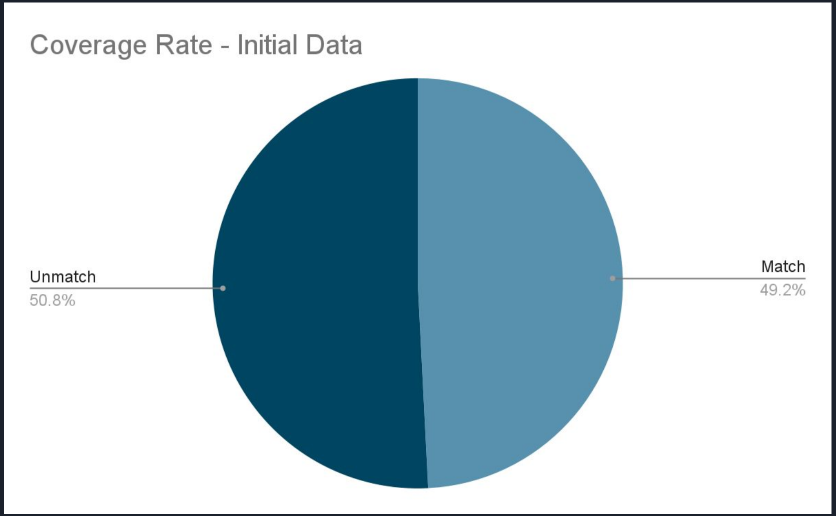
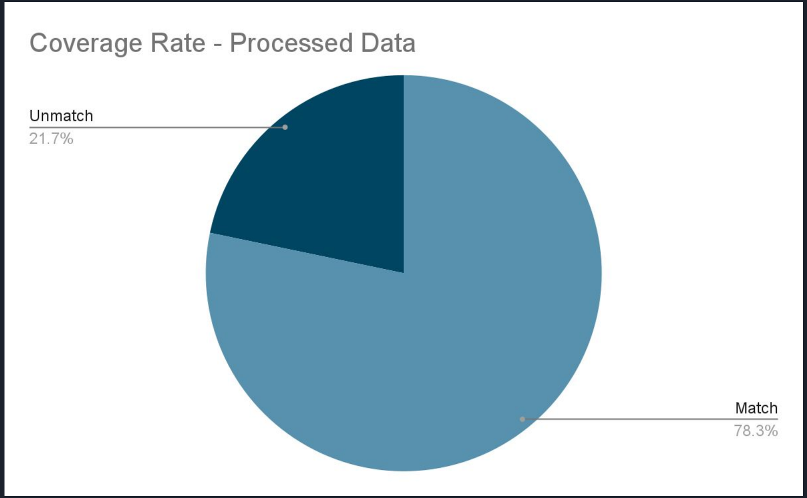
Part 1:
Data Quality

By: Chanpreet Singh

Result


Coverage Rate - Initial Data
Unmatch
50.8%
Match
49.2%


Coverage Rate - Processed Data
Unmatch
21.7%
Match
78.3%

Analysis

- Data address match was increased by 29.08%, from initall 49.19% to 78.27%.
- Current Method
    - Remove unnecessary info such as Unit #
    - Fix format by adding ‘,’ and adding space between street number and name
    - Updating addresses with 3 or less than 3 words different.
    - Updating missing cities where state and zip code match.
    - Updating any incorrect information by comparing corresponding matched row values.
- Further Possible Improvements
    - Using python libraries such as ‘FuzzyWuzzy’ or ‘usaddress’
        - Tried using them but they take too long to work on large datasets. Could use them by moving datasets to SQL or other forms and faster the processing with higher processing computer.
    - Could train a machine learning model to match the addresses such as Random Forest.
        - Didn't think that the problem needed machine learning model solution.
    - Could use external address match directory such as Canada Post AdressesComplete API in Canada or Google Map API. This will increase address accuracy overall with most updated data.