Data Integration - 2. Assignment
Group M

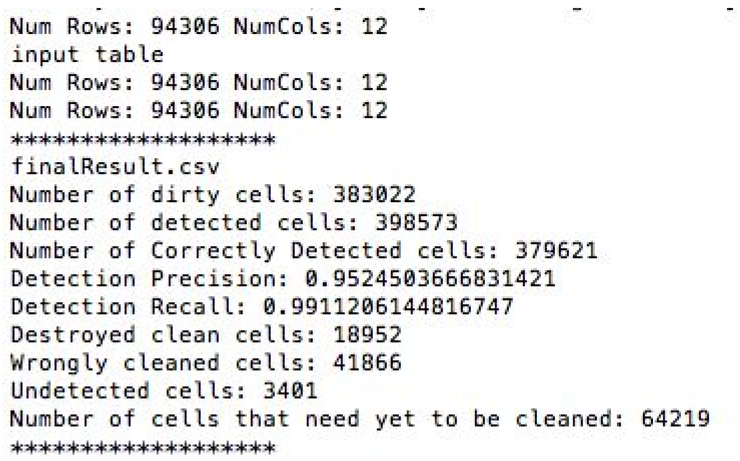
The cleaning job is done using Java and with the help of USPS API.

Cleaning strategy:

- All non-numerical characters are converted to upper case.
- DOB, PO BOX and the POCityStateZip are ignored, as per assignment.
- FirstName, MiddleName, LastName are also ignored, because there is no way to validate people’s names.
- SSN cleaning :
    1. Remove all characters except numbers (incl. whitespaces).
    2. The number of digits should be between 8 and 10, but there is no way to correct them if that is not the case.
- ZIP cleaning :
    1. Remove all non numerical characters from Zip column.
    2. Check the length, if they are not of length 5, check the zip code in POCityStateZip (the zip codes might be different, but the city and state they refer to should be the same).
- City and State cleaning :
    1. Use a mapping between state name and state code to check if it is possible to convert any state name to state code.
    2. Check in the Zip column if there is any non-numerical characters (A-Z), if there is, it is possible that it refers to the state name/code. If that is the case, copy the possible state name/code to the state column, and any value in the state column to the city column (the common error is that if the city name consists of two or more words, only the first word would be put in the city column, and the rest in state column, e.g “Palm Springs” would be “Palm” in city column, and “Springs” in state column).
    3. Also check if the city name is accidentally put in the last part of the address column.
- Address validation :
    1. Check the validity of the complete address (address, city, state, zip) using the USPS API. If the address is valid, the API will return the complete address (corrected) along with its city, state code and zip code.

Time spent on task: ~1 week

The final performance score is as follows:



```
Num Rows: 94306 NumCols: 12
input table
Num Rows: 94306 NumCols: 12
Num Rows: 94306 NumCols: 12
*******************
finalResult.csv
Number of dirty cells: 383022
Number of detected cells: 398573
Number of Correctly Detected cells: 379621
Detection Precision: 0.9524503666831421
Detection Recall: 0.9911206144816747
Destroyed clean cells: 18952
Wrongly cleaned cells: 41866
Undetected cells: 3401
Number of cells that need yet to be cleaned: 64219
*******************
```