Data Integration

Assignment 2: Data Cleaning

Group BB

Student Name	Student ID	Course of Study
Shady Alkhouri	389561	Computer Science, Master
Karel Smejkal	388638	Computer Science, Master
Ayham Kanhoush	391136	Computer Science, Master

Task 1 - Error Detection:

1. To find error in cell we need to check if the cell is formated as is specified in the task. Probably most obvious errors are not capitalized characters or wrong length of **State**, **ZIP** or **SSN** column. More difficult is to check if City column contains real city name. For this we need external API. We use python module uszipcodes https://pypi.python.org/pypi/uszipcode and method `search._find_city(cityName, best_match=True)`. This method get the most similar city name to the given city.
2. ----------Error Detection Results----------
   Number of detected cells: 415611 (Number of changed values)
   Number of Correctly Detected cells: 383691 (cell was correctly identified as an error)
   Detection precision: 0.923197412965 (ratio of correctly detected cells over all detected cells)
   Detection recall: 0.988573298362 (ratio of correctly detected cells over all erroneous cells in the data)
   Detection F1: 0.954767542119

Task 2 - Error Correction:

1. To change the cell to its correct value we again used the python module uszipcodes https://pypi.python.org/pypi/uszipcode with the methods to lookup for **city** and **state** if we know the correct **zipcode**, or use city/state to lookup for zipcode and/or city and state if the zipcode is not in right format, which is not really accurate because one city can have more zipcodes or even there can be two cities with same name in different state. To change **SSN** to the correct value is probably impossible because every person has the first 3 numbers in the SSN based on the stated they were born in and the rest is unknown to us, because there is no general rule to make this number.
2. ----------Error Correction Results----------
   Destroyed clean cells: 31920 (cell was correct but has been transformed into a wrong value)
   Wrongly cleaned cells: 31080 (cell was wrong but the cleaning was also not correct)
   Undetected cells: 4435 (cell was erroneous but was not touched)
   Number of cells that need yet to be cleaned: 67435 (sum of the 3 cell types above)
   Correction precision = 0.848415946642 (ratio of correctly corrected cells over all changed cells)
   Correction recall = 0.908496209994 (ratio of correctly corrected cells over all erroneous cells in the data)

Note: The module is really slow to lookup for best matched names for cities(it can take over 2 hours even on better laptops to finish) so I also added the corrected .csv file and jupyter notebook file that is faster to run than the .py file. And also we were using python3, so we needed to modify print method of web_client.py