The 800 Pound Gorilla in the Corner: Data Integration
Assignment 2: Data Cleaning

June 2018

GROUP L

Henrik Dichmann	358361
Fiona Wille	376585
Ariane Ziehn	358027

1 General information

Given DB schemas:

InputDS(RecID; FirstName; MiddleName; LastName; Address; City; State; ZIP; (POBox; POCityStateZip); SSN;(DOB))

Data Quality Constraints

1. All alphabetical characters in all columns should be capitalized.
2. Address data should be compatible to the standard in https://tools.usps. com/-go/ZipLookupAction!input.action.
3. The State column should contain the correct two character US state code.
4. City column should contain real city names.
5. ZIP column should be formatted as a 5 digit value. (col 7; row 6)
6. SSN column should contain an 8-10 digit value. (col 10; row 9)
7. Note that, you do not need to fix the DOB, POBox, and POCityStateZip columns.

2 Task: Error Detection

Task: Here, we want to only detect data errors. To mark a cell as data error, you just need to change its value into something else.

2.1 Which of the mentioned data quality constraints can help you to detect data errors? How?

1. All alphabetical characters in all columns should be capitalized.
    - Yes, this can be detected with a regular expression, matching any lower case character.
2. Address data should be compatible to the standard in https://tools.usps. com/-go/ZipLookupAction!input.action.
    - Could be used for checking if Zip code and address matches. We skipped using this service because our precision was higher than 90% even without this tool.
3. The State column should contain the correct two character US state code.
    - Yes, can be detected by checking the String length. This is not 100% accurate as only syntax can be tested but not semantic. A mapper containing all possible state codes helps to higher the detection rate of wrong character combinations. For 100% accuracy the content of the columns City, State and ZIP would need to be considered to test this column.
4. City column should contain real city names.
    - No, only by the usage of a mapper containing all real city names a syntax test could be applied, but there are probably to many cities to create a useful map.
5. ZIP column should be formatted as a 5 digit value.
    - Yes, String length and regular expression to check if only digits are contained.
6. SSN column should contain an 8-10 digit value.
    - Yes, String length and regular expression to check if only digits are contained

2.2 Report your best error detection precision, recall, and F1.

- Error detection precision: 93.87%
- Recall: 93.86%
- F1: 93.86%

3 Task: Error Correction

Task: Here, we want to not only detect data errors, but also automatically correct them to the true values. Therefore, it is not enough to just mark a cell as data error, you also need to update it to the correct value.

3.1 Which of the mentioned data quality constraints can help you to correct data errors? How?

1. All alphabetical characters in all columns should be capitalized.

   - Yes, this can be detected with a regular expression, matching any lower case character. If a lower case letter is contained it will be replaced with its matching upper case letter.

2. Address data should be compatible to the standard in https://tools.usps. com/-go/ZipLookupAction!input.action.

   - Could be used for checking if Zip code and address matches. We skipped using this service because it would significantly impact the performance of the algorithm in a negative way. Since correction recall rates were as high as 77 % even without using the service, we decided to skip this part. Future implementations could include the use of the web-service.

3. The State column should contain the correct two character US state code.

   - Yes, can be detected by checking the String length. This is not 100% accurate as only syntax can be tested but not semantic. A mapper containing all possible state codes helps to higher the detection rate of wrong character combinations. For 100% accuracy the content of the columns City, State and ZIP would need to be considered to test this column.

4. City column should contain real city names.

   - No, only by the usage of a mapper containing all real city names a syntax test could be applied with the help of a database containing all real city names. In our implementation we just correct lower case to upper case characters and do not perform any spell checking or comparison with a database.

5. ZIP column should be formatted as a 5 digit value.

   - Yes, firstly delete spaces and letters then check String length and regular expression to check if only digits are contained.

6. SSN column should contain an 8-10 digit value.

   - Yes, firstly delete spaces and letters then check String length and regular expression to check if only digits are contained.

3.2 Report your best error correction precision, recall, and F1.

- Error correction precision: 77.40%
- Recall: 77.39%
- F1: 77.40%