Big Data - Spring 2024
Data Cleaning Lab

Consider the tax dataset (file 'tax100.csv' accompanying this assignment), which contains information about US tax records. Each record describes an individual's address and tax information with 14 attributes: first and last name, area code, mobile phone number, city, state, zip code, marital status, having children status, salary, tax rate, tax exemption amount if single, married, or having children.

Your task is to write scripts that identify and return data errors in the Tax dataset.

No need to use super specialized libraries – a simple script in Python or any other programming language should suffice. For some questions, you may also consider using a database management system (DBMS).

Pattern Violations

The phone number must adhere to the pattern YYY-YYYY, where Y represents a numerical digit. Write a script to identify and return all rows that do not adhere to this pattern

Duplicate Detection

For the simplicity of this Lab, we can assume duplicates in the Tax dataset as any two records that share identical or closely similar values for both First Name and Last Name. Let us employ Edit Distance as a measure of similarity, with a threshold set to at most 1. For example, "John Wonder" and "Jon Wander" would be considered duplicates.

A python library for Edit Distance: https://pypi.org/project/editdistance/

Let us explore two variations of the solution:

1. Implementing an iterative (nested loop) version.
2. A variant version of 1 that uses a blocking approach, where records are grouped by their City and State attributes.

Compare the performance (runtime) of these two solutions. To better understand the performance impact, use the larger version of the Tax dataset (tax1k.csv file).

Constraint Violation

Assuming that the following constraints apply to the Tax dataset:
- The Zip code determines the City.
- If two persons live in the same state, the one earning a lower salary should have a lower tax rate.

Write scripts or queries that return the pairs of tuples that violate these rules.

Suppose you only need to determine whether the rules hold in the Tax dataset without showing the violations. Is there an alternative query/solution to the problem? Do you expect that to be faster?

Constraint Discovery

You aim to provide users of the Tax dataset with rules for data cleaning purposes.

1. Write a script that identifies all unique column combinations of size one (such as the column TID) that exist in the Tax dataset.

2. Additionally, write a script to discover all functional dependencies (FD) of the form A -> B that hold in the Tax dataset. For simplicity, assume that both A and B represent single attributes. An example of an FD that holds in the Tax dataset is Zip -> State.