

CS 171/CSCI E-64: Visualization

Homework 1, Problem 4: Cleaning Data

James Goodspeed – jgoodsp@fas.harvard.edu

The data in the `massachusetts-crime.csv` and `massachusetts-unemployment.csv` files were cleaned up in the following ways:

1. Any cities that had a state abbreviation had that abbreviation removed to keep the data consistent. This step was done manually in Google Refine as there were only a few changes to make.
2. The population column was formatted to remove the comma so that the data could be interpreted as a number instead of text. This change was made using Google Refine:
 - a. Create a filter of the records that had a comma in the population column
 - b. Use the following transformation to remove the comma:
`value.replace("\",", "")`
3. The Unemployment data had additional towns that the Crime data did not. The towns that did not exist in the crime csv were removed from the merged data using Google fusion tables.
4. All of the number columns except population were changed to have one decimal place to keep the data consistent. This change was made with a combination of `sed` and `awk`.
5. The data was sorted on the town name alphabetically to make the data easier to read using the Unix `sort` command.