Rob Acheson
racheson@fas.harvard.edu
CS-171 - Homework 1
2/7/2013

cleanup_justifications.pdf

After installing Google refine, I imported the two data sets and started to examine them. I began with *massachusetts-crime.csv*. I took a very linear approach to working with this data set and moved left to right, starting with City. The following is a description of actions performed on each row.

> **City:** Using a text filter I searched for *", MA"* and since there were only a few results, I manually edited the fields. Not the most efficient way, but I was just getting started.
>
> **Population:** Realizing that there were far too many results to manually edit, I started looking into text transformations and the Google Refine Expression Language (GREL). I attempted to use *value.toNumber()*, but that failed because of commas. So I first removed the unwanted characters, combining into one action; *toNumber(replaceChars(value, ",", "")).*
>
> **Remaining Fields:** These looked pretty good as-is. However, I noticed that a few were empty, which could lead to problems. Row-by-row, I used a numeric facet quickly isolate row values that were blank. Then on the group I could apply a text transformation and set all values to *0* in one action. As an added bonus, numeric facets revealed the range of the data across all fields, which could come in very handy for finding random or inaccurate data values.

Happy with the data, I exported a copy of the data to csv and started on *Massachusetts-unemployment.csv.* I took a very similar approach to the number fields and again used numeric faceting to isolate problem rows; the *City* column was a different story. With too many instances of *", MA"* in the row to manually edit, I opted to find the index of the first comma and end the string there using *substring(value, 0, indexOf(value, ","))* Again finding the data properly sanitized, I exported a copy as csv.

To merge the tables, I used *Fusion Tables*. After importing both documents, so that they both were available in *Google Drive*, I used the file merge tool on *massachusetts-crime-csv-clean*. Because this file had fewer rows (276) compared with *massachusetts-unemployment-csv-clean (351)* the non-matching 75 rows were lost in the merge. At first I though this was a mistake and quickly repeated the process in the opposite order and verified that I could get all of data. However, if the purpose of the exercise were to build a data set to correlate unemployment and crime, there would be no need to keep the 75 unique rows. Seems like I stumbled across a simple technique to keep only relevant data.