Lab 10

In this lab you will be using OpenRefine to wrangle data.

Introduction

OpenRefine is an open source tool for working with bad data. You can download OpenRefine using this link: http://openrefine.org/

To get an introduction to OpenRefine you can either read the documentation, or follow this tutorial: http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

We will be using two data sets one from with earthquake data and one with customer complaint data.

The first data set is the eq2015 data set which data about earthquakes of magnitude 3 or more during the first 6 months of 2015. You can fine a data attribute glossary here

You can download the data set here.

The second data set contains customer complaints, you can download that data set here.

Please answer the following questions by using OpenRefine.

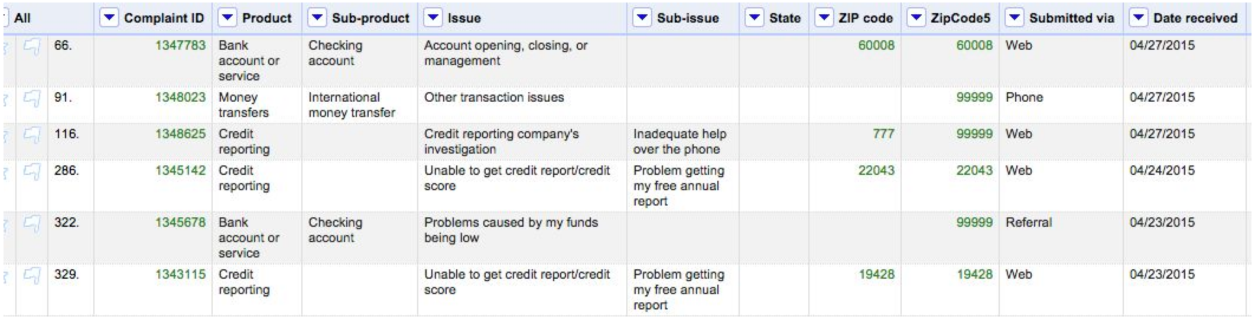
Wrangling the Customer Complaints Data

- A1: How many rows are missing value in the state column? Explain how you came up with the number?

- A2: How many rows with missing zip codes do you have?

- A3: Clean up the zip code column. Create a new column called “ZipCode5” with all zip codes that contains 5 digits preserved. All other rows should have the zip code 99999. You should have the same type for all cells in the created column.

Example of result:



All	Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	ZipCode5	Submitted via	Date received
66.	1347783	Bank account or service	Checking account	Account opening, closing, or management			60008	60008	Web	04/27/2015
91.	1348023	Money transfers	International money transfer	Other transaction issues				99999	Phone	04/27/2015
116.	1348625	Credit reporting		Credit reporting company's investigation	Inadequate help over the phone		777	99999	Web	04/27/2015
286.	1345142	Credit reporting		Unable to get credit report/credit score	Problem getting my free annual report		22043	22043	Web	04/24/2015
322.	1345678	Bank account or service	Checking account	Problems caused by my funds being low				99999	Referral	04/23/2015
329.	1343115	Credit reporting		Unable to get credit report/credit score	Problem getting my free annual report		19428	19428	Web	04/23/2015

- A4: If you consider all zip codes less than 99999 valid zip codes. How many valid and invalid zip codes do you have respectively.

Cleaning up eq2015 Data.

- A5: For column “nst” fill in missing values.

- A6: Clean up the place column so that it has state or country name depending on what is in the text.

- A7: From the column “updated” extract the Date without time into a new column called “eventdate”

- A8: Run cluster en edit on “location” column. Run nearest neighbor and levenshtein distance. Answer the following questions:
    - Does it make sense to merge detected values?
    - Why or why not?

- A9: Try to do nearest neighbor clustering on “place’ column.
    - What happens?
    - Explain why it is happening.

Help information and URLs

- You can find a useful OpenRefine cheat sheet here:
    - http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf
- Some additional documents on OpenRefine can be found here:
    - http://davidhuynh.net/spaces/nicar2011/tutorial.pdf
    - http://schoolofdata.org/handbook/recipes/cleaning-data-with-refine/
- Reference to the GREL expression language:
    - https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language

- Data Sets:
  - Earthquake data glossary: http://earthquake.usgs.gov/earthquakes/feed/v1.0/glossary.php#net

  - Earthquake Data: https://github.com/UC-Berkeley-I-School/w205-labs-exercises/blob/master/lab_10/dataset/eq2015.csv

  - Consumer Complaints Data:https://github.com/UC-Berkeley-I-School/w205-labs-exercises/blob/master/lab_10/dataset/Consumer_Complaints.csv