MIDS W205

|  |  |  |  |
| --- | --- | --- | --- |
| **Lab #** | 10 | **Lab Title** | Cleaning Data With OpenRefine |
| **Related Module(s)** | 10 | **Goal** | Get hands on introduction to OpenRefine |
| **Last Updated** | 9/24/15 | **Expected duration** | 60 minutes |

Cleaning Data with OpenRefine.

## Introduction

OpenRefine is an open source tool for working with bad data. You can download OpenRefine

here To get an introduction to OpenRefine you can either read the documentation, 1 or follow t his 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 download the data set here . You can fine a data attribute glossary 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?

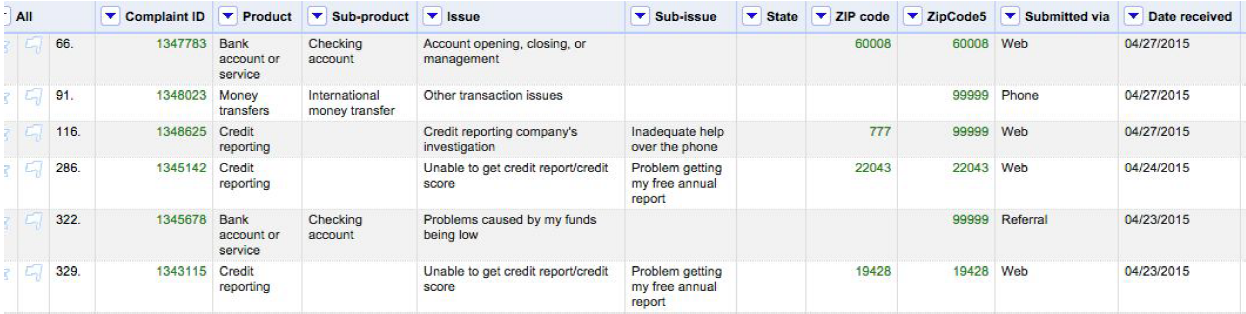
● A2: How many rows with missing zip codes do you have?

● 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:



● 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

● OpenRefine Links:

○ Download: http://openrefine.org/

○ Cheat sheet: http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf

○ Tutorial: http://enipedia.tudelft.nl/wiki/OpenRefine\_Tutorial

● Some additional documents on OpenRefine can be found here:

○ Tutorial 2: http://davidhuynh.net/spaces/nicar2011/tutorial.pdf

○ Tutorial 3: http://schoolofdata.org/handbook/recipes/cleaningdatawithrefine/

● Reference to the GREL expression language:

○ GREL:

https://github.com/OpenRefine/OpenRefine/wiki/GeneralRefineExpressionLang

uage

● Data Sets:

○ Earthquake data glossary:

http://earthquake.usgs.gov/earthquakes/feed/v1.0/glossary.php#net