# Module 6: Homework 1: Assignment: Data Prep with OpenRefine

## Download and Installation

* Go to <https://openrefine.org/download>
* Download the .zip file, and extract it into a folder where you wish to store program files to C:\Program Files\OpenRefine.
* Install Java https://adoptium.net/temurin/releases/?os=windows
* Open the application by right-click on openrefine.exe or refine.bat.

A screenshot of a computer

Description automatically generated

## OpenRefine GUI

* OpenRefine automatically open to a browser window, or I can manually navigate to the application's URL (using <http://127.0.0.1:3333>).

A screenshot of a computer

Description automatically generated

## Create a New Project & Upload Data

* Click > Create Project tab on the left side
* To select the data sources, chose files > Uploaded file, then click Open
* Click next and write your Project name > click Create project

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Check States with Text Facets

* Select the Column: “State” > click Text facet. I could see all the states in the file as well as their distribution of unique values.

A screenshot of a computer

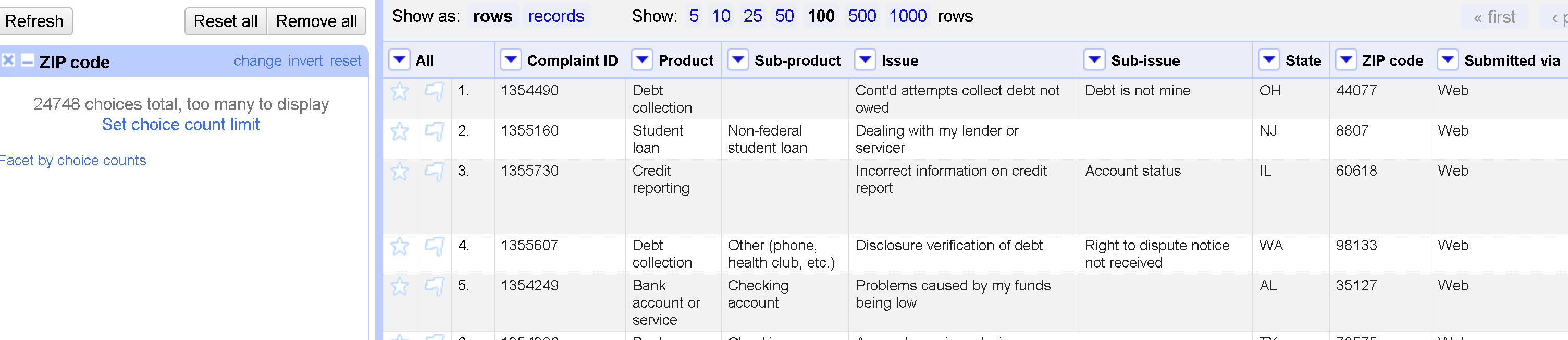
Description automatically generated

A screenshot of a computer

Description automatically generated

## Wrangling/Munging – Transforming Data: Check Zip code

First I was ablet to check the txt facet for the zip code. The result shows 24748 total zipcodes. And it shows all the values are string filed. I follow the step to convert the data to Number.



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Cleansing/Wrangling Data: Handling Missing Zip codes

* In this step, I try to fill in the missing values
* Step Add new column using the new Name “ZipCode5” and using Expression if(value.length() > 4, value, 99999)
* A screenshot of a computer

  Description automatically generated

A screenshot of a computer

Description automatically generated

## Data Set eq2015: Cleansing & Wrangling Data:

* I opened a new browser and created a new project with the EQ2015 data set.

A screenshot of a project

Description automatically generated

## Started Wrangling/Munging Data: Transforming the Place Column

* Extract transforming the Place Column > Click drop-down mean of Place Column > Add Column based on this column. Enter value.split(‘,’)[1] for expression to split substring based on coma.

A screenshot of a computer

Description automatically generated

## Data Set eq2015: Cleansing & Wrangling Data – Location Column

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Exploring/Cleansing/Wrangling Data Set by Clustering

* Clustering: - this feature is designed to identify groups of cell values that may represent the same thing,

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Cleansing/Wrangling Data: Editing Cells

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Cleansing/Wrangling Data: Manually Editing Cell

- Go to Location, Click edit, and update the name.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated