ADTA 5240 – DATA PREP WITH OPENREFINE
PART I: Cleaning & Wrangling Data
Yog Chaudhary

Data wrangling, cleaning, or munging is about fixing and shaping data to answer questions. Analytics people spend most of their time (80%) fixing data and the rest (20%) analyzing it. It makes data easier to use by changing it in smart ways, like fixing a whole column of data or mixing columns together. If data is collected or shown poorly, we need to wrangle it. Data from people can be mistaken, and data from websites might not be set up well for analysis. This process helps fix these issues and makes the data ready for analysis.

Remove Irrelevant Values: The elimination of pointless values is among the most fundamental data-cleaning techniques used in data mining. You should erase all unnecessary data from your machine as soon as possible. You don't need any information that is pointless or unimportant. The context of your problem might not allow for it. This kind of data frequently does not fit the situation you are attempting to evaluate.

Get rid of Duplicate values: Getting rid of duplicate values in your data is important because they just take up space and don't help your analysis. These repeats can happen for many reasons, like combining data from different places, mistakes when entering data, or clicking "enter" too many times on a form. Duplicate values are a common issue in datasets. Finding and removing these duplicates, a process known as de-duplication is a key part of cleaning up your data. This helps make your dataset more accurate and easier to work with, saving you time in the long run.

Uniformity of Language: Uniformity of language in data is crucial, especially when using Natural Language Processing (NLP) models for analysis. These models usually work with just one language at a time. If your data has multiple languages, it could cause problems because the system might not understand all of them. So, when cleaning your data, make sure that all the information is in the same language. This ensures that your analysis is accurate and that the NLP models can process the data correctly.

Avoid Typos: Typos are common and can be fixed using different tools and methods to ensure all words are spelled correctly. It's important to correct them because models see different spellings and cases as different values. For example, "George" with an uppercase "G" is not the same as "George" with a lowercase "g". Fixing these mistakes helps in making your data more accurate and reliable for analysis.

Remember, the goal of data cleaning is to improve the quality and reliability of your data, making it easier to analyze and derive insights.

URL: https://www.upgrad.com/blog/data-cleaning-techniques/

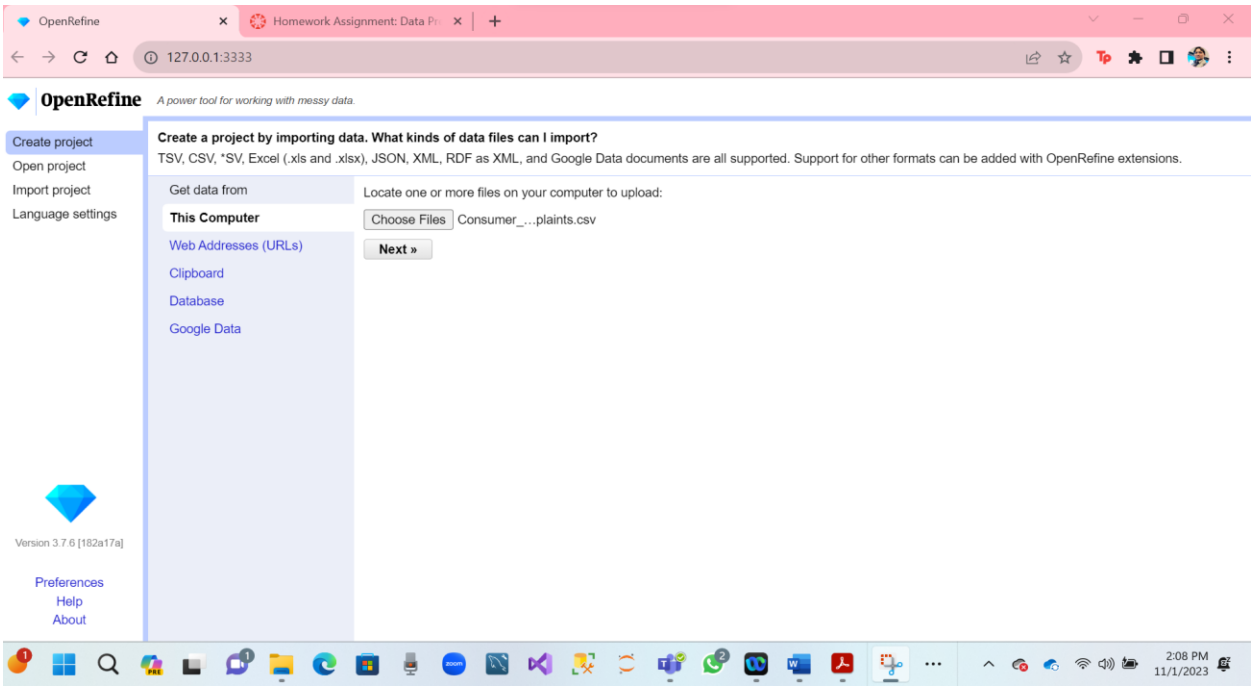
PART II: Exploring, Cleaning, Wrangling Datasets

Data Prep with Open Refine

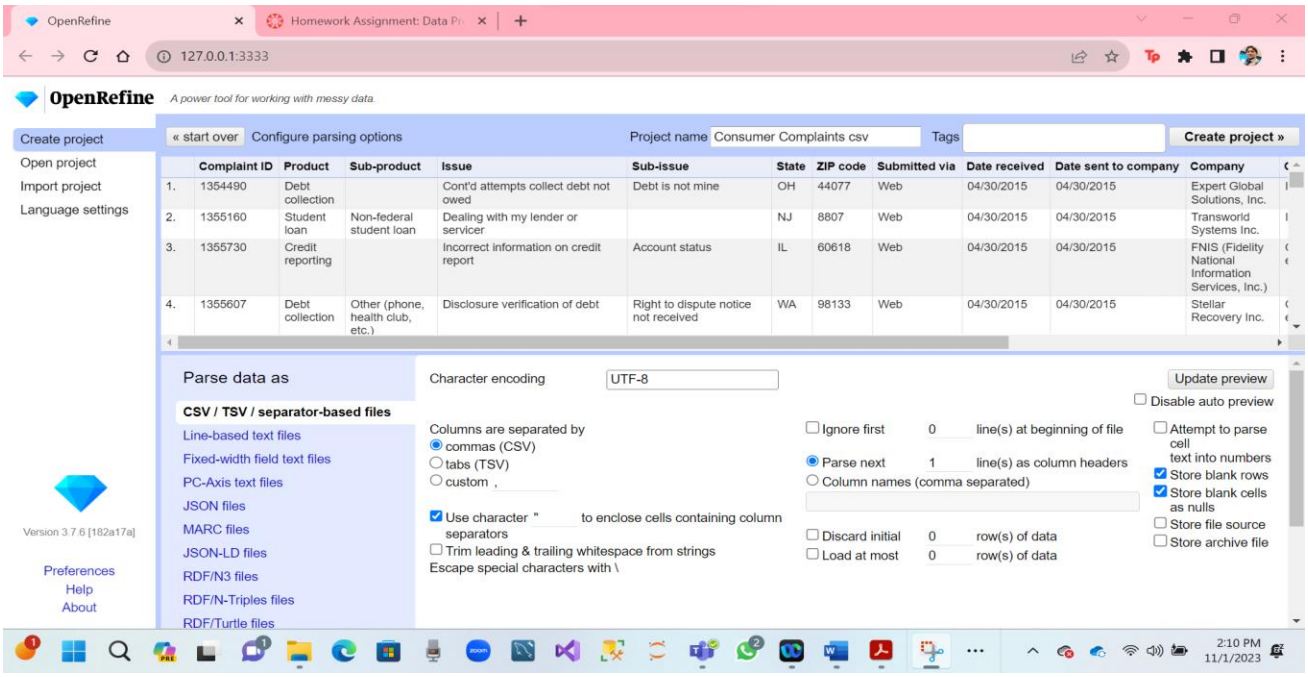
I have downloaded OpenRefine by using this link (https://openrefine.org/docs/manual/installing)
After downloading, I was taken to the OpenRefine page and showed a web-browser GUI pops up. (home screen: http://127.0.0.1:3333/)

Creating a new project and uploading data.
Consumer_Complaints.csv: cleansing & Wrangling Data

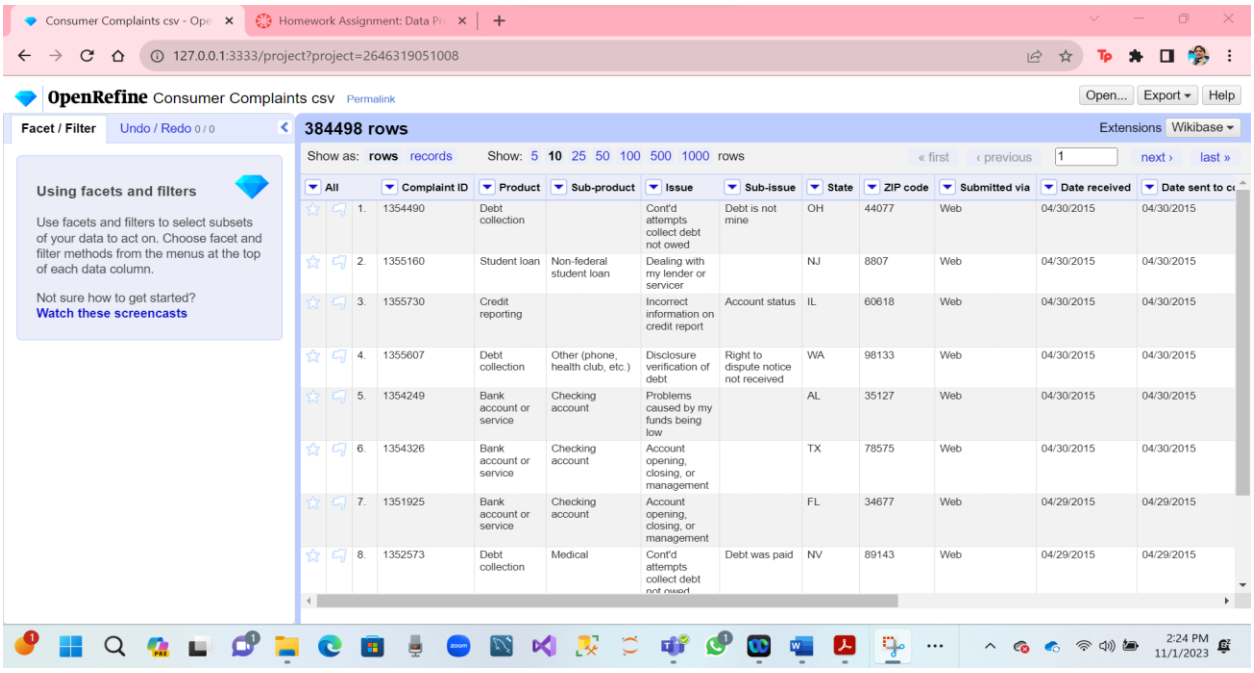
- Hare, I clicked on choose files.
- Then I browsed for Consumer_Complaints.csv.
- I clicked on next.



- Hence the Consumer_Complaints.csv dataset uploaded.
- After this, I was taken to the below page.



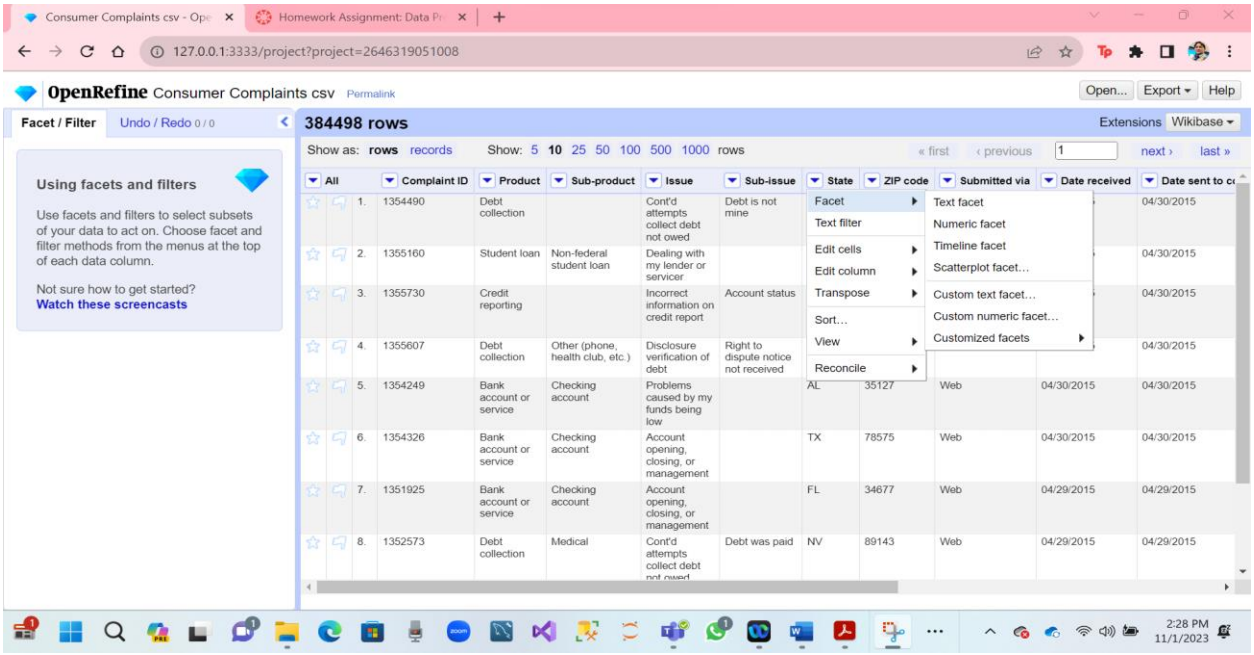
- Here I left defaults as checked i.e.,
- OpenRefine detected that the data file is in CSV format.
- Line (Row) is a header line.
- Quotation marks are used to enclose cells containing column separators.
- Stone blank rows.
- Store blank cells as nulls.
- Then I clicked on Create a project.
- After that I was taken to the below page.



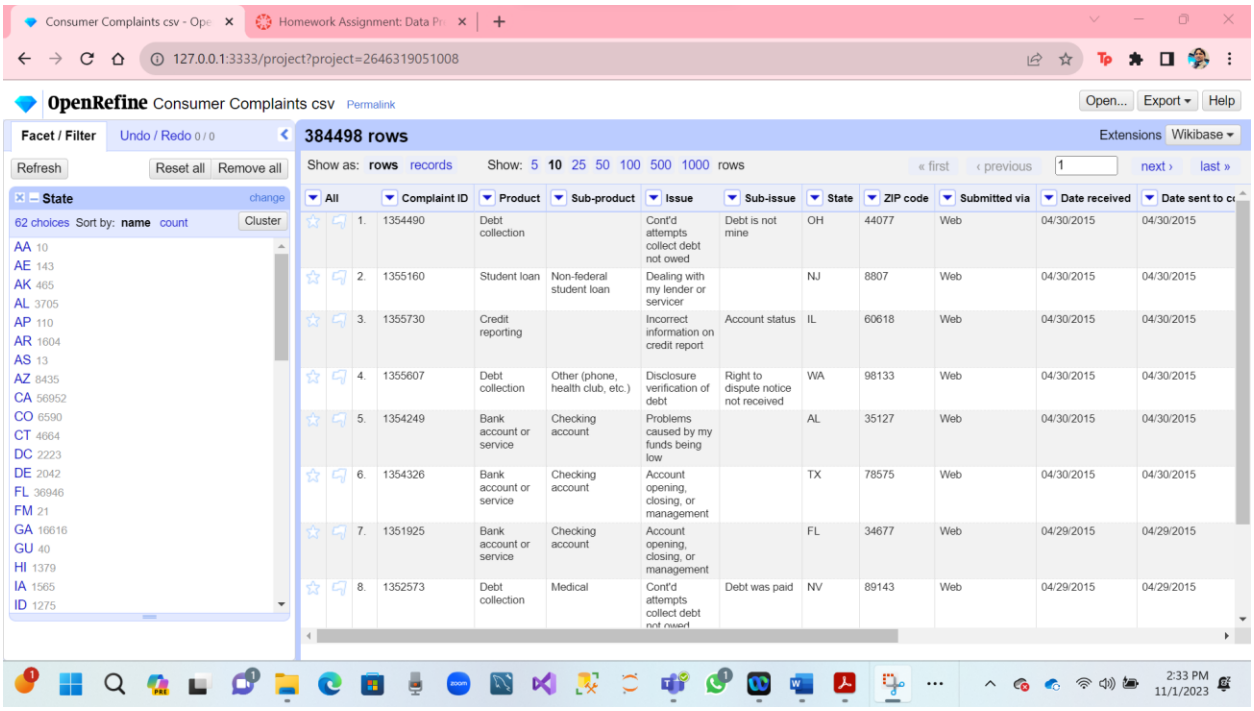
- Here I have seen 384,498 rows that are loaded.
- After I clicked on “50” where OpenRefine shows 50 rows at once.
- If we want to see other rows(records) we can click next.

2. Check states with Text Facets:

- Here we are analyzing the data: we will use “Text Facet to display the number of occurrences of each unique value in a column.
- It is like a filter in Excel.



- Here I have clicked on the drop-down menu arrow of state.
- Then I clicked on Facet and Text Facet, after clicking on that I was taken to the below page.

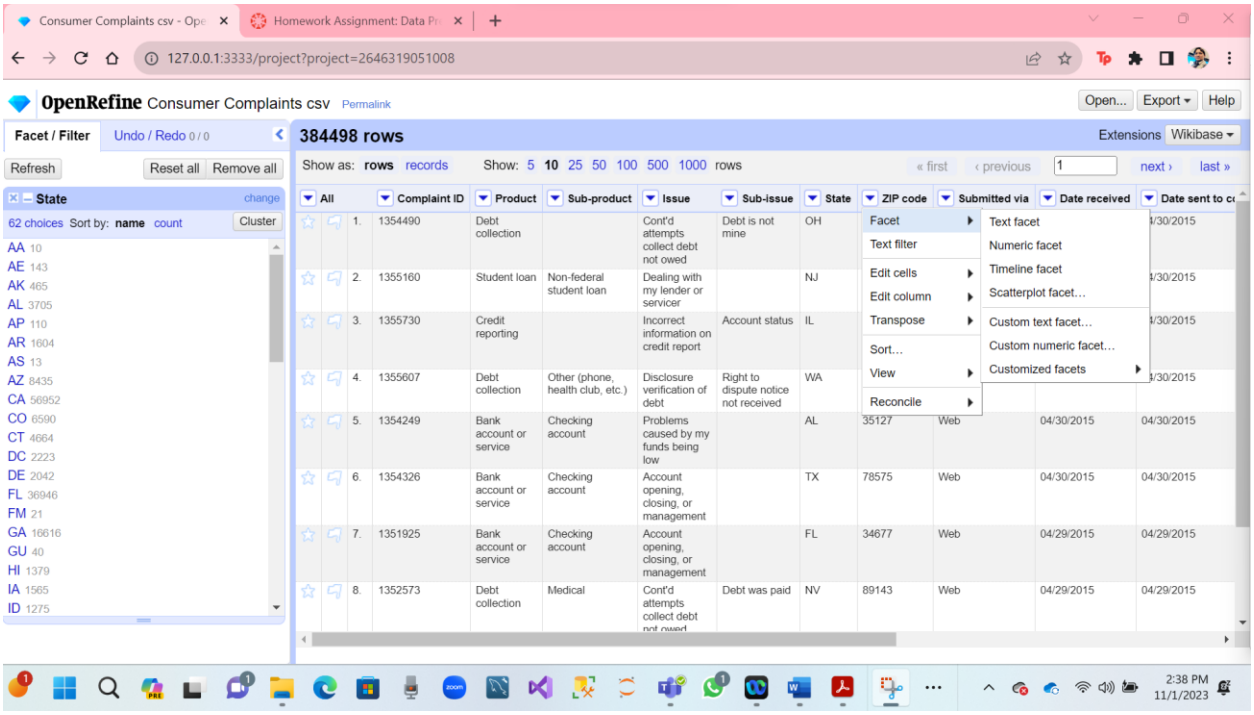


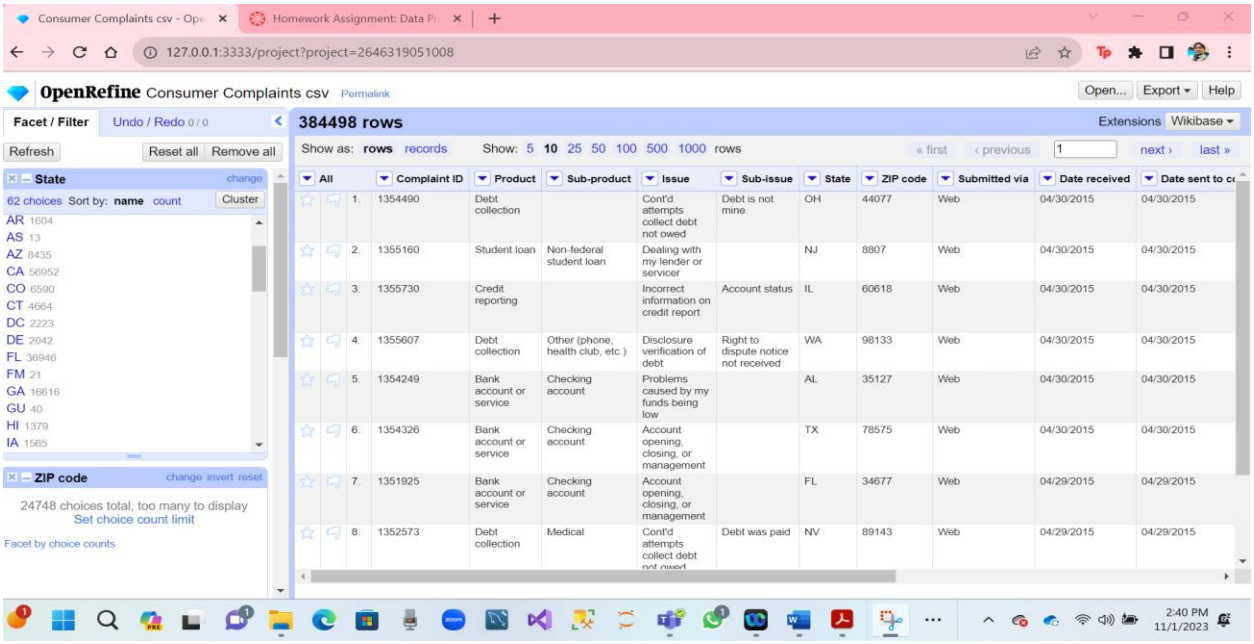
- Here I have seen 62 states (51 states and other US territories) that are listed in the panel.
- Then I clicked on remove all to clear the result panel.

3. Wrangling/Munging – Transforming Data: Checking Zip code.

a. Text Facet for Zip code:

- I clicked 0 drop-down menu arrow of the zip code.
- Then click on Facet and Text Facet.
- After clicking on Text Facet, I was taken to the below page.

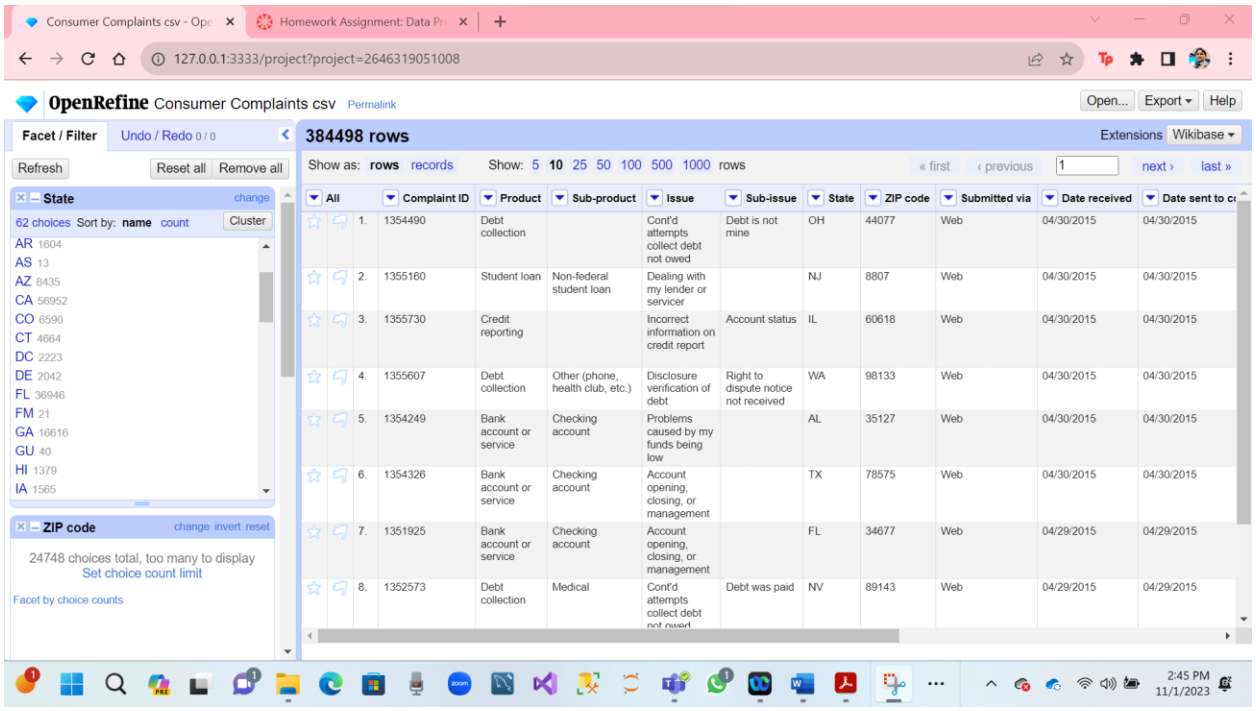




- Here the result shows there are 24,748 zip codes in the dataset.
- Some of the zip codes are missing i.e., blank, like in row 57.

b. Numeric Facet:

- I clicked on the drop-down menu arrow of the Zip code.
- Then click on Facet and Numeric Facet.
- After clicking on that I was taken to the below screenshots.

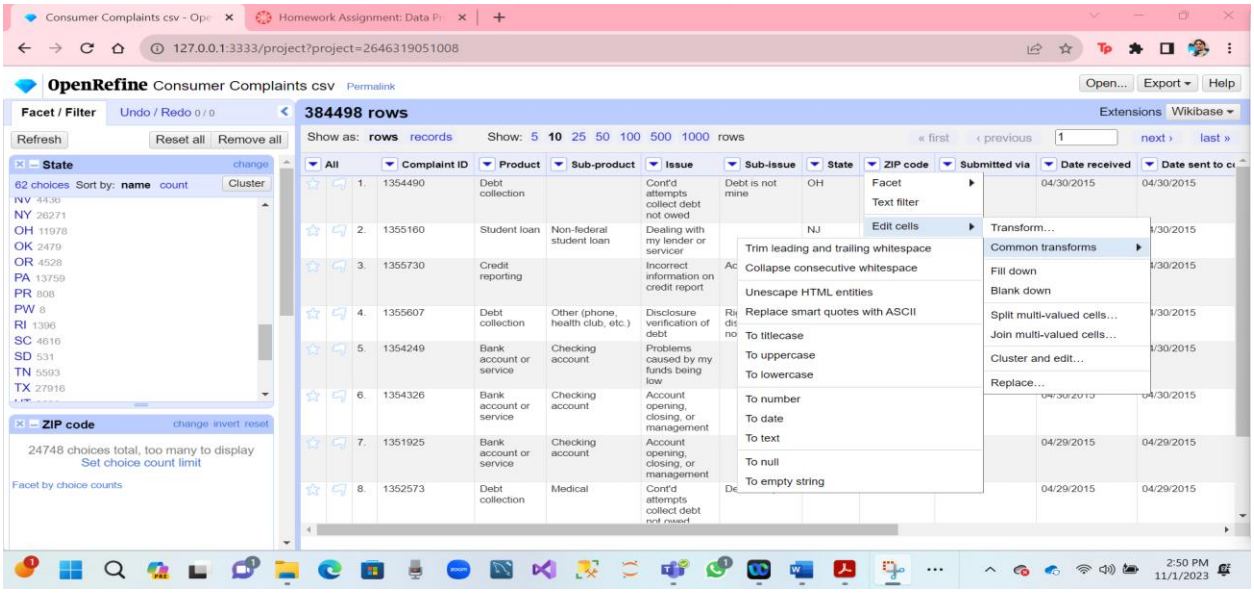


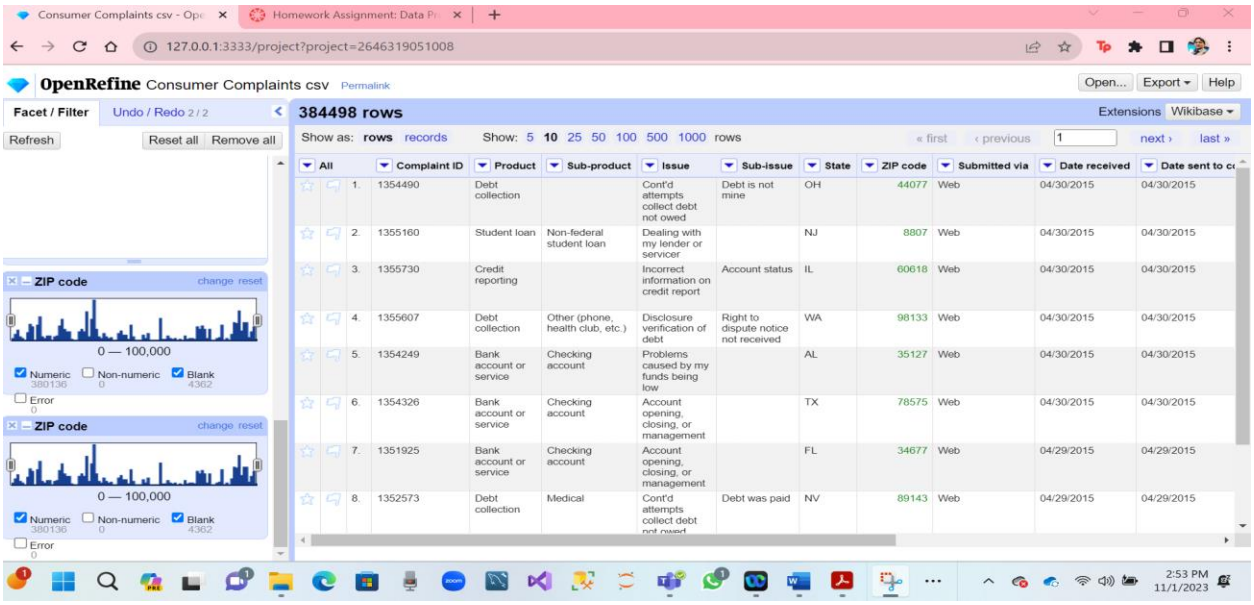
- Here there is no numeric zip code in the data set, which means all of them are included as text values (24748 zip codes).

C. Transform zip code from Text to Numeric (Wrangling Data).

It is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- I clicked on the drop-down menu arrow of the zip code.
- Then click to select Edit cells.
- Then click on Common transforms and To number.
- After clicking on that I was taken to the below page.



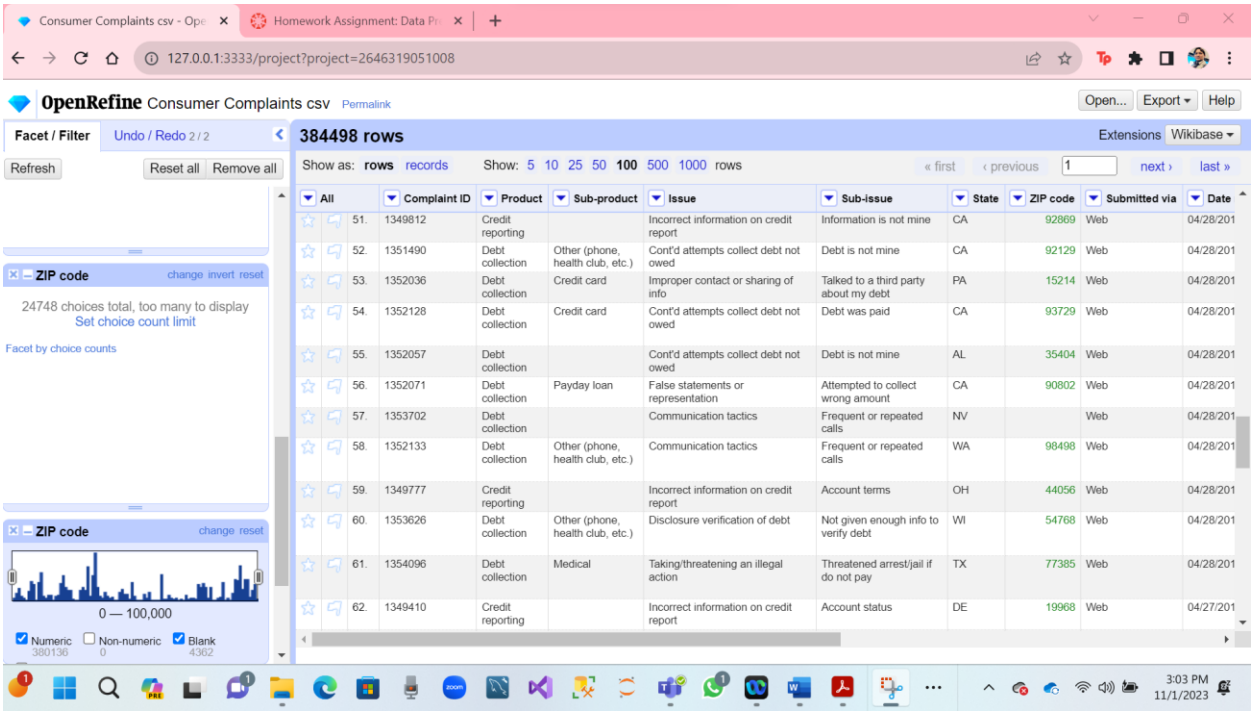


- In the above screenshot, results show 380136 zip code text values have transformed into numeric values and results show 4362 zip codes are missing.

4. Cleansing/Wrangling Data: Handling Missing Zip Codes

a. Fill down the missing value:

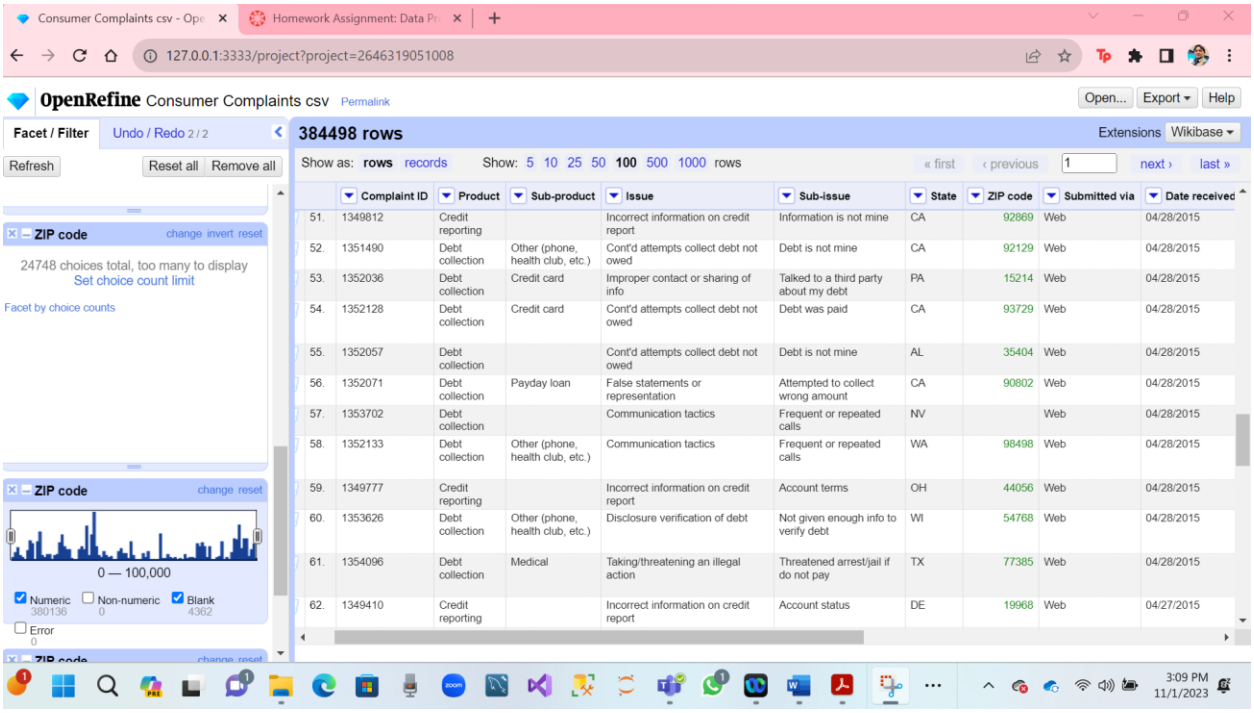
- One way of filling in missing values is taking the previous value and using that to set subsequent empty cells.
- Now looking at row 57, the value of the zip code is missing.
- I clicked on the drop-down menu arrow of the zip code.
- Then click on edit cells and fill down.
- After clicking on that I was taken to the below page.



- The results show there are no missing zip codes now which means all missing values are filled.
- Now when we look at row 57, here the zip code value is 90082, which was filled from row 56.
- In row 57, the state is NV, not CA.
- Can't be filled down to handle the missing values of zip code.
- Must rescind the wrangling data.

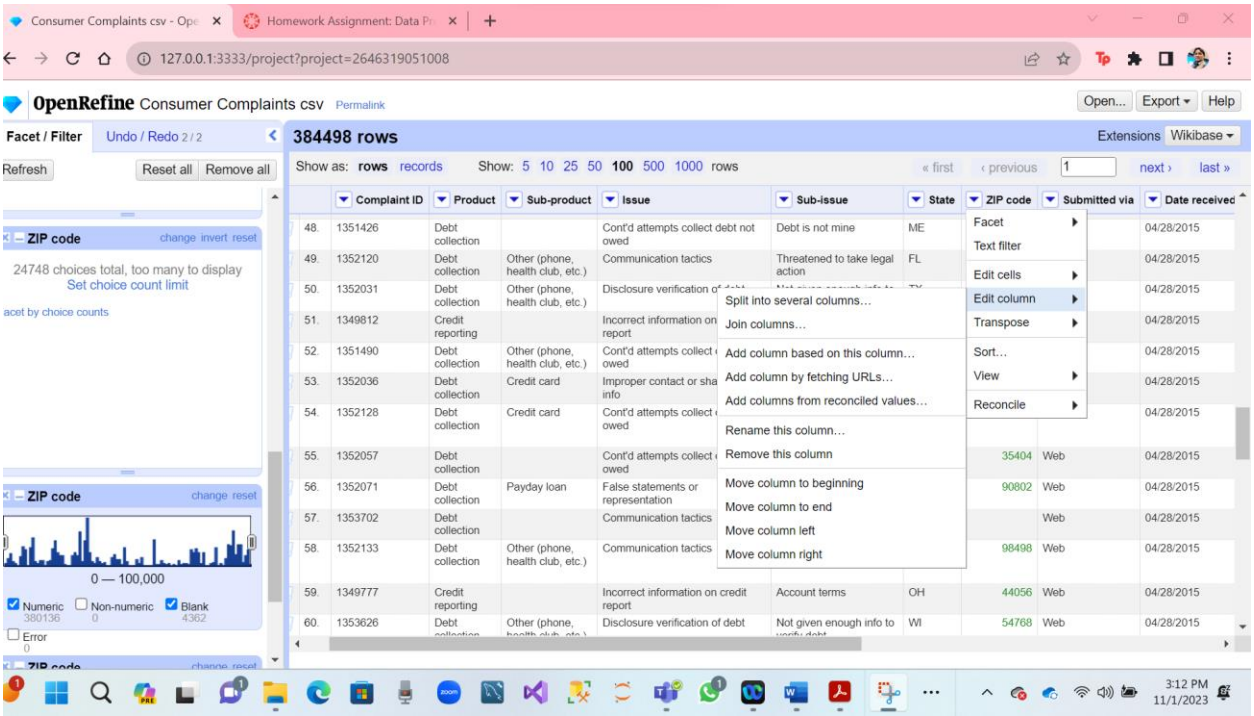
Rescind the wrangling data:

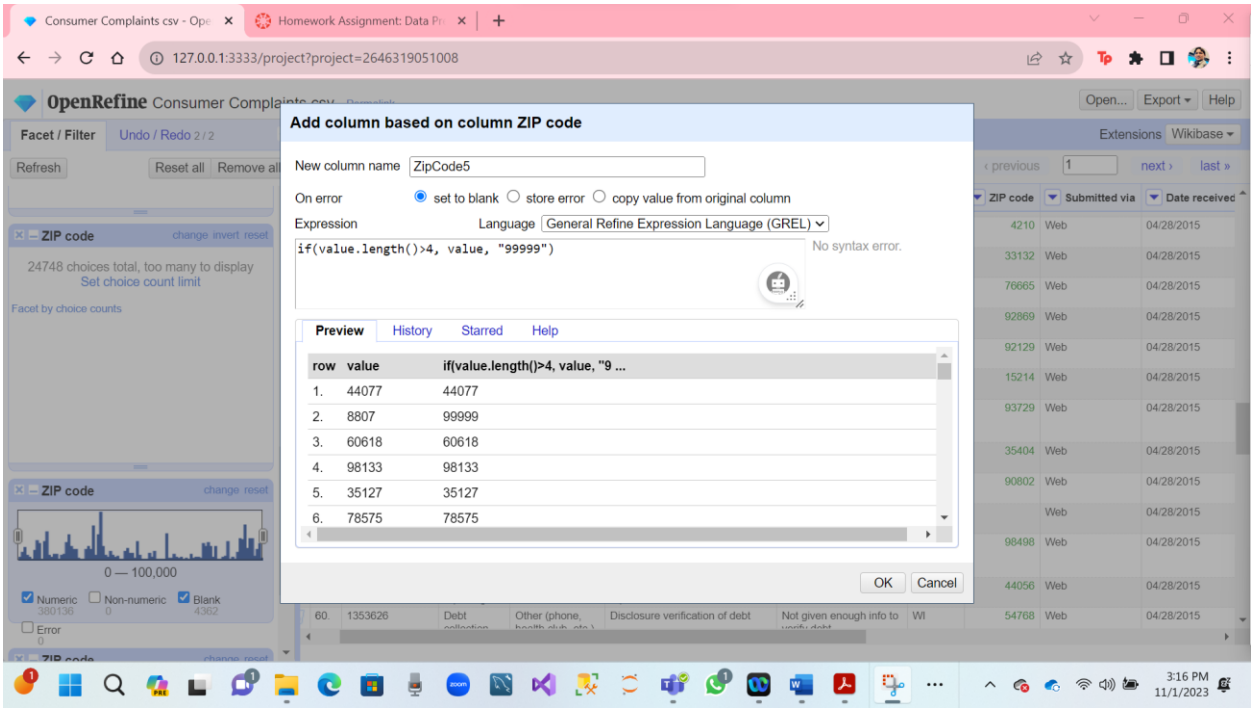
- OpenRefine has very advanced UNDO functionalities.
- For undoing the previous step, I clicked on undo as shown in the below page.
- UNDO lists 3 levels of activities 0,1,2 which can be rolled back.
- the last level 'fill down 4362 cells in column zip code' was highlighted.
- Then I clicked on "Text Transform on 380136." as the latest activity.
- By clicking on this the zip code in row 57 was undone as shown below.



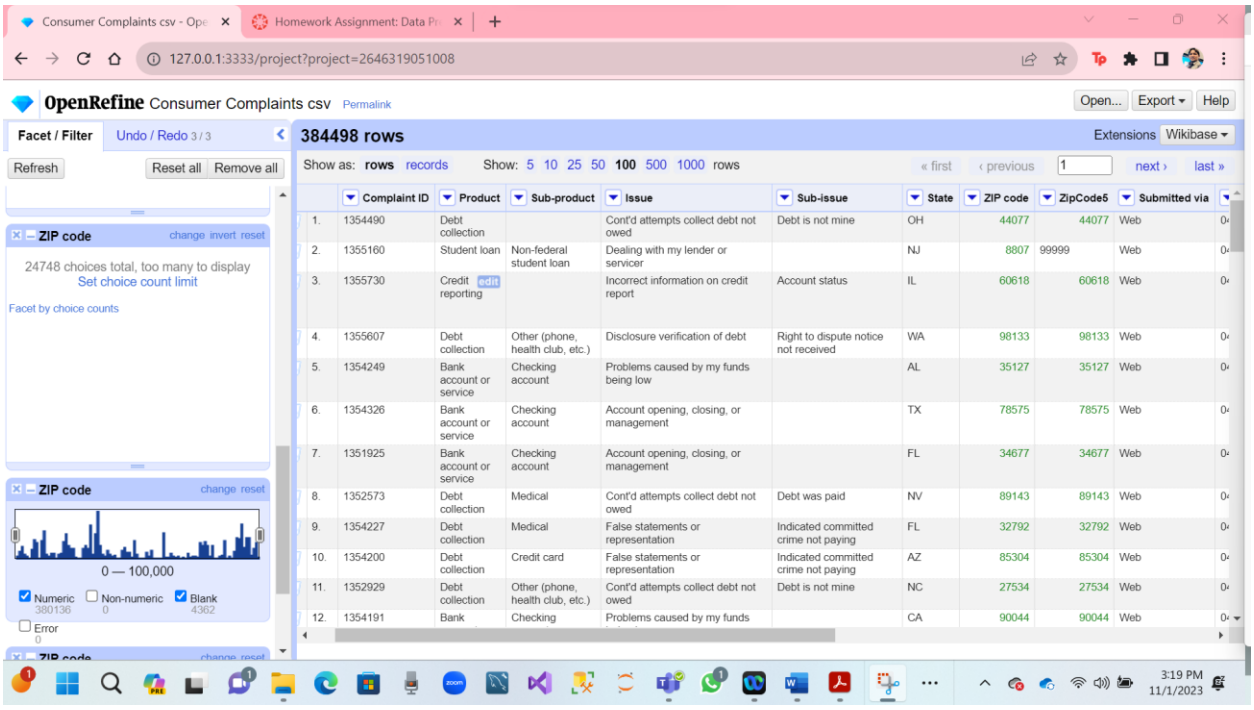
b. Handling Missing Zip codes: Create a New Column.

- I clicked on the drop-down menu arrow of the zip code.
- Then click on the edit column and add a column based on this column.
- After clicking on that I was taken to the below page.

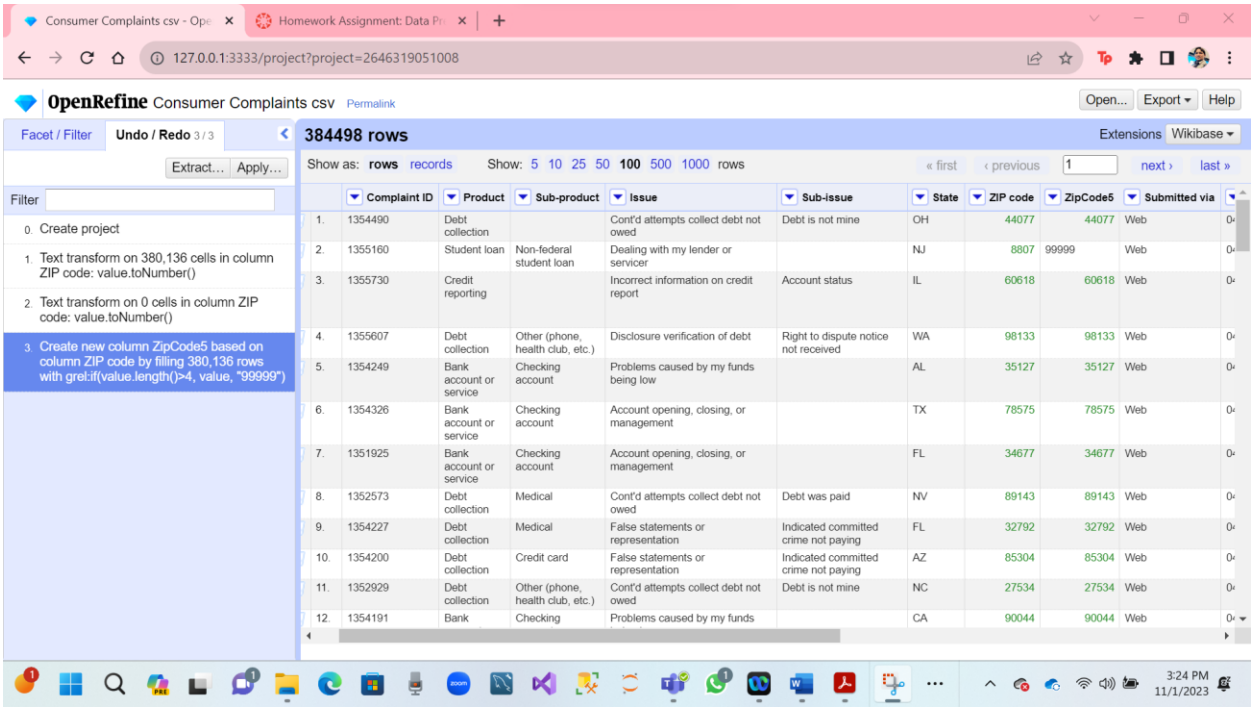




- For performing cleansing/wrangling/transforming data we use a scripting language.
- Here I have given the column name ZipCode5 and entered.
- If(value.length()>4, value, “99999”).
- Then I clicked on OK.
- After clicking on I was redirected to the below page.



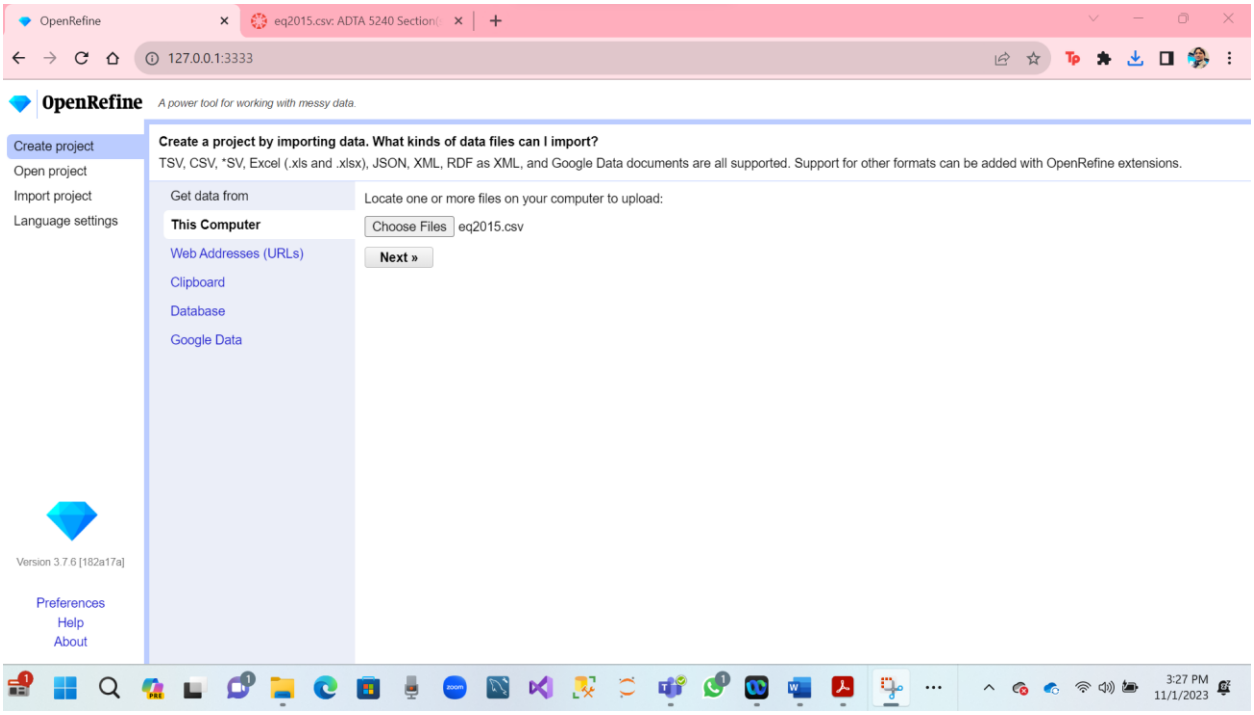
- Here the results show a new column zip code5 has been created.
- When I looked at row 2 the original zip code has 4 digits and a new one has 5 digits (99999), however, it justified in left. But this value is not numeric.
- Then I must undo the operation and start over.
- Again, I clicked on the drop-down menu arrow of the zip code.
- Then click on the edit column and add a column based on this column.
- After clicking on that I was taken to the below page.
- Here I have given column name ZipCode5 and
- If(value.length()>4, value, 99999).
- Then I click on OK.
- After clicking on OK I was taken to the below page.



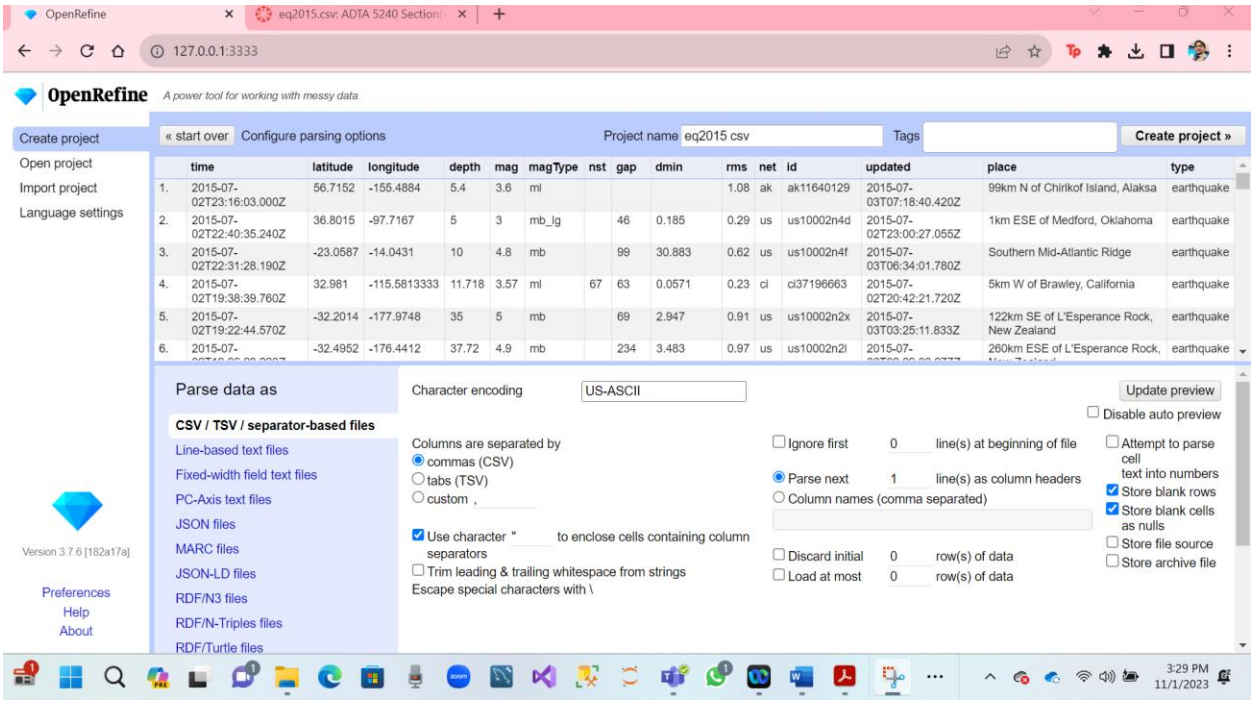
- Here the results show a new column zip code5 has been created.
- When I looked at row 2 the original zip code has 4 digits and a new one has 5 digits(99999), however, it justified to right.

❖ **Data set eq2015: cleansing & Wrangling Data: Getting started.**

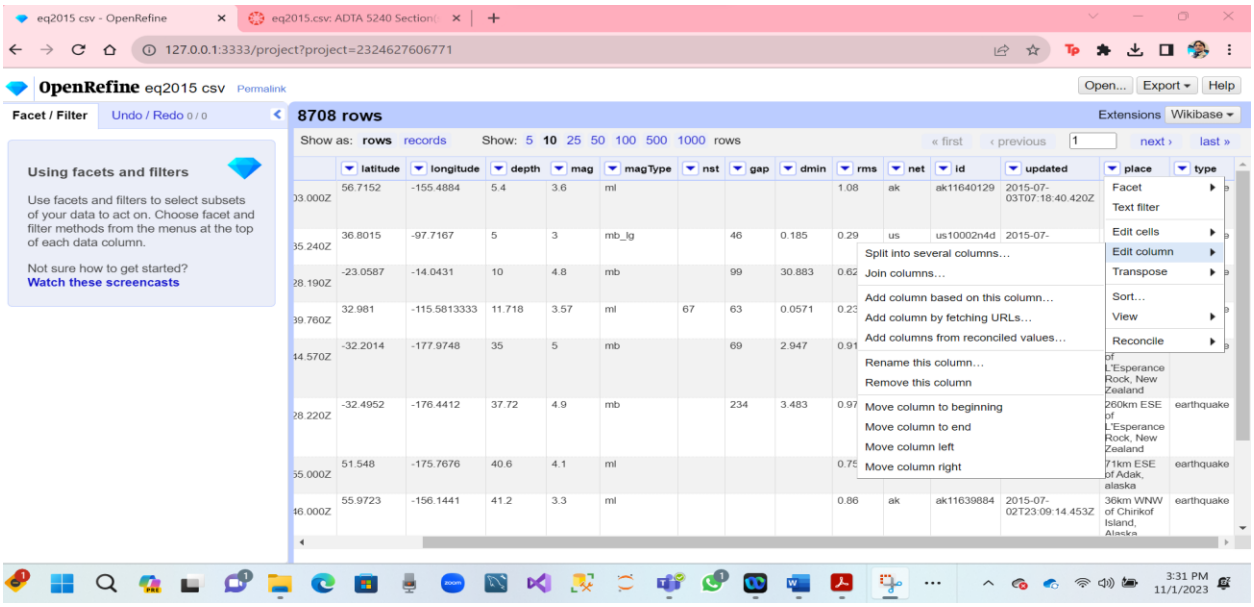
- Here I clicked on Choose files.
- Then I browsed for eq2015.csv.
- Then I clicked on next



- Hence the eq2015.csv dataset has been uploaded.
- After this I was taken to the below page.



- Then I left defaults as checked and clicked on create project.
- By clicking on create project I was taken to the below page.

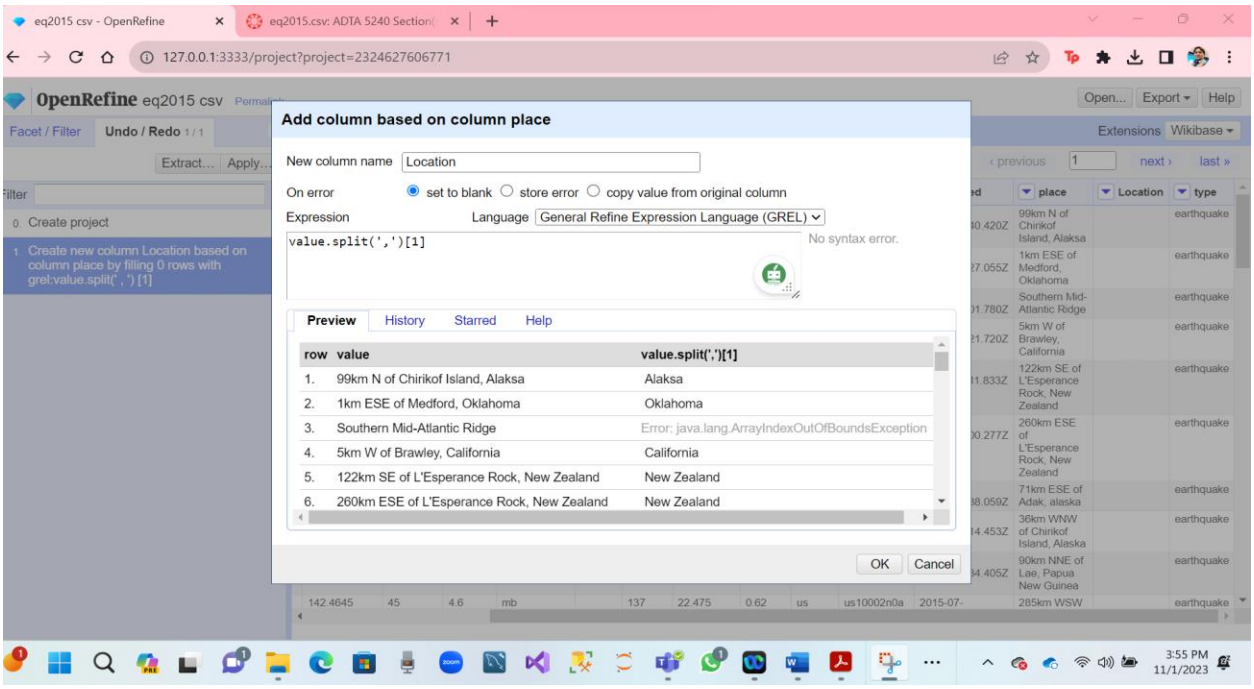


- Then I clicked on 50, and it showed me 50 rows.
- In the above page “nst” column was missing few values.

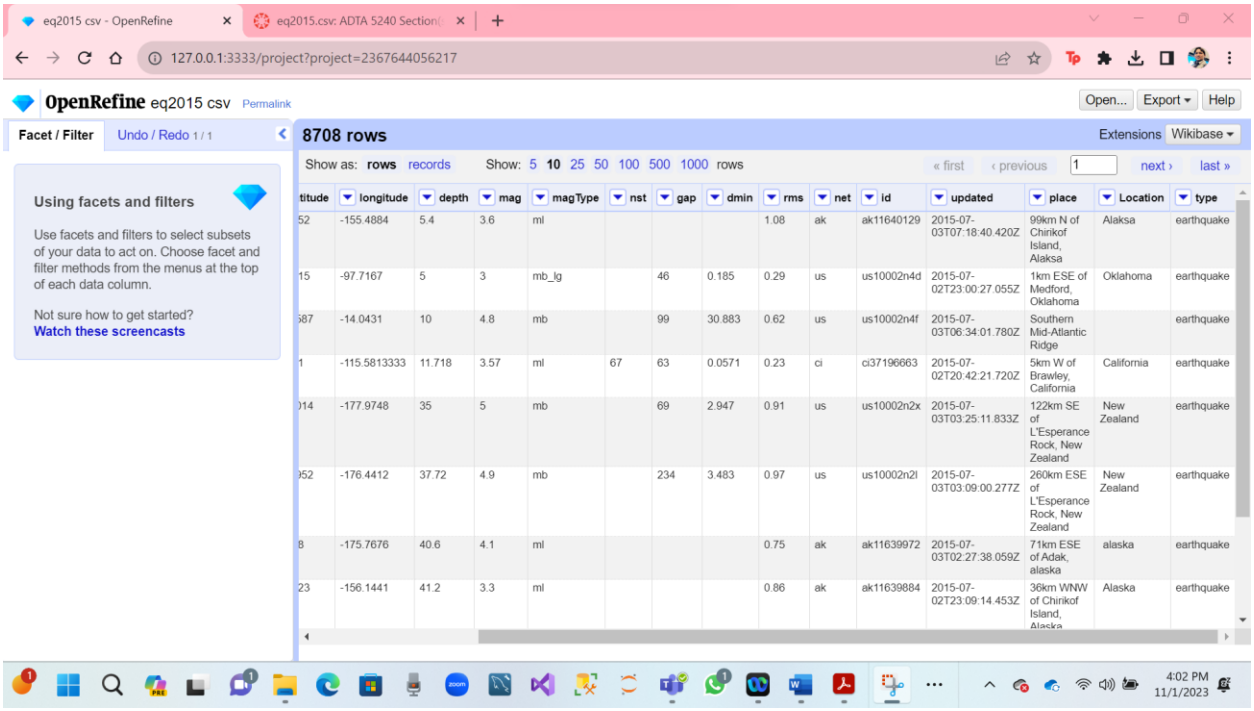
1. Wrangling/Munging Data: Transforming the place column:

Transforming the data:

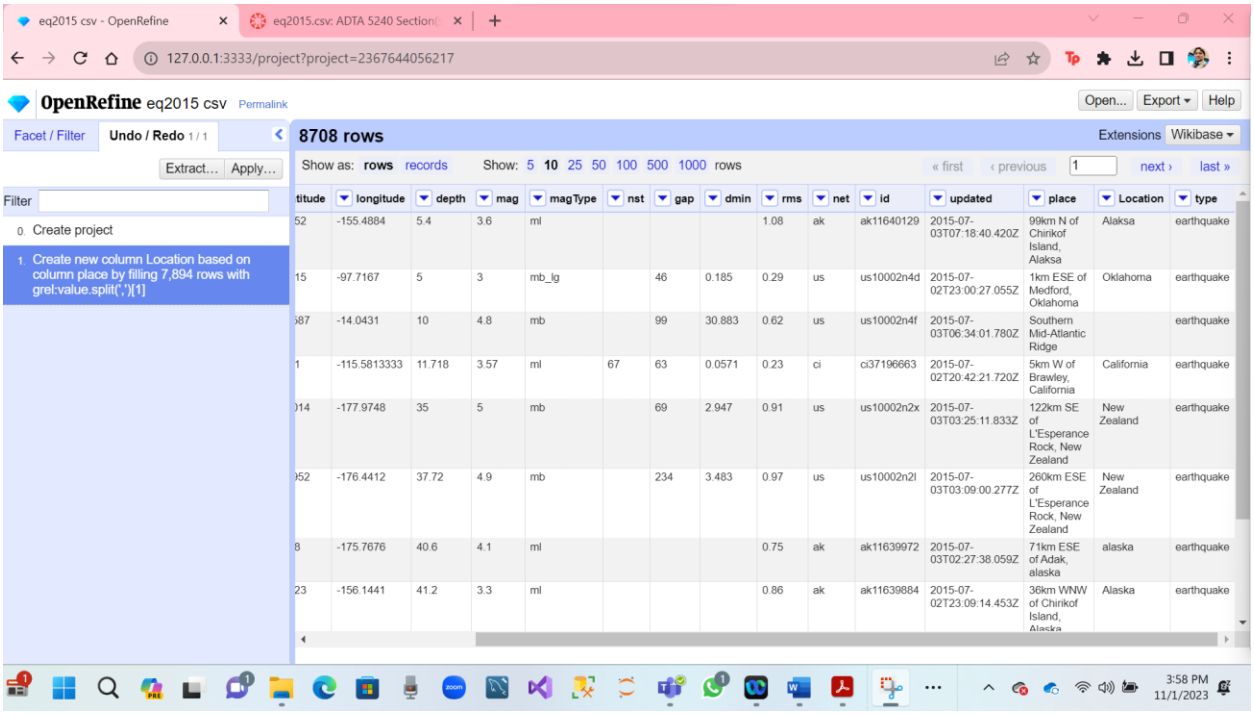
- Here I am going to extract to approximate area from the “place” column.
- I clicked on the drop-down menu arrow of place.
- Then click on the edit column and add a column based on this column.
- I was redirected to the below page.



- I have given “location” for the name of a new column.
- I have given,
- Value. split(‘,’) [1] in expression
- Then click on OK.
- We use “,” for splitting the two substrings.
- After clicking on ok I have taken to the below page

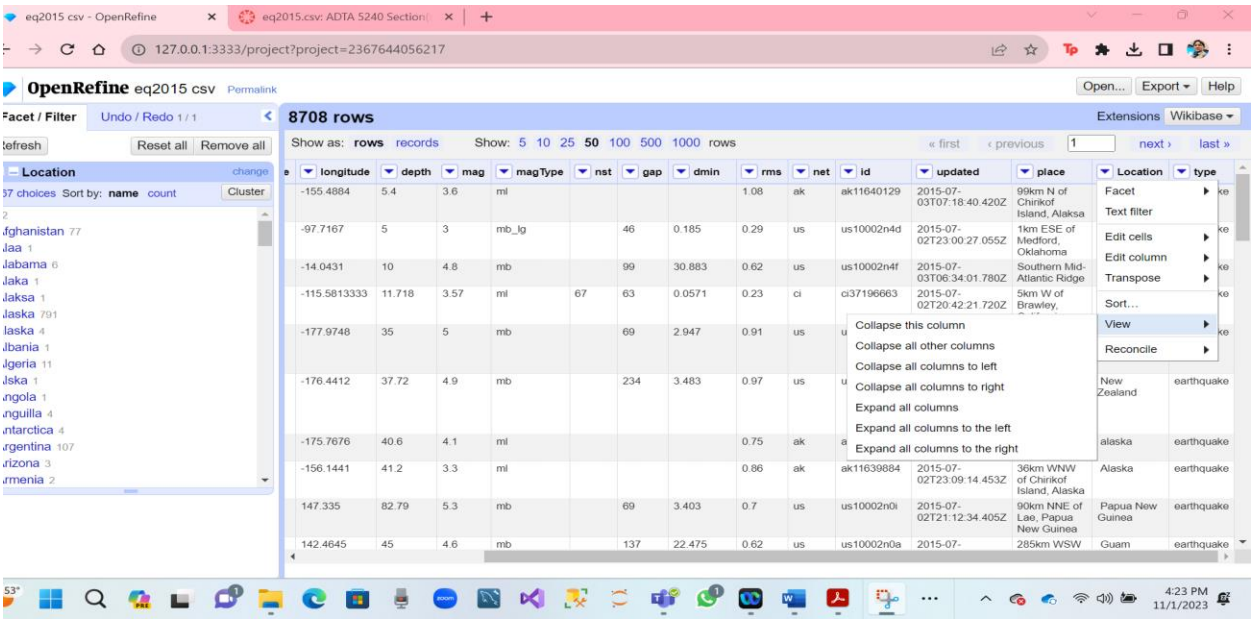


- In this we can see a problem in row 3, which means not all the values have 2 strings separated by”,”.
- So, to handle this we need to change the expression.
- Then click on edit column and add column based on this column.
- Then I have entered,
- If(value.split(“,”).length()<2, “offshore”, value.split(“,”[1]) for expression.
- After clicking I was redirected back to the below page.



2. Data set eq2015: Cleansing & Wrangling Data – Location Column:

- Here we have multiple strings like Alaska, but some of them appear to be misspelled.
- When we look at row 14(Alaska), row 18 (Alaska), etc.
- It was the same for California, so before we need to remedy them, we need to conduct further research to understand the data.
- For this I clicked on the drop-down menu arrow of location.
- Then click on Facets and Text facets.



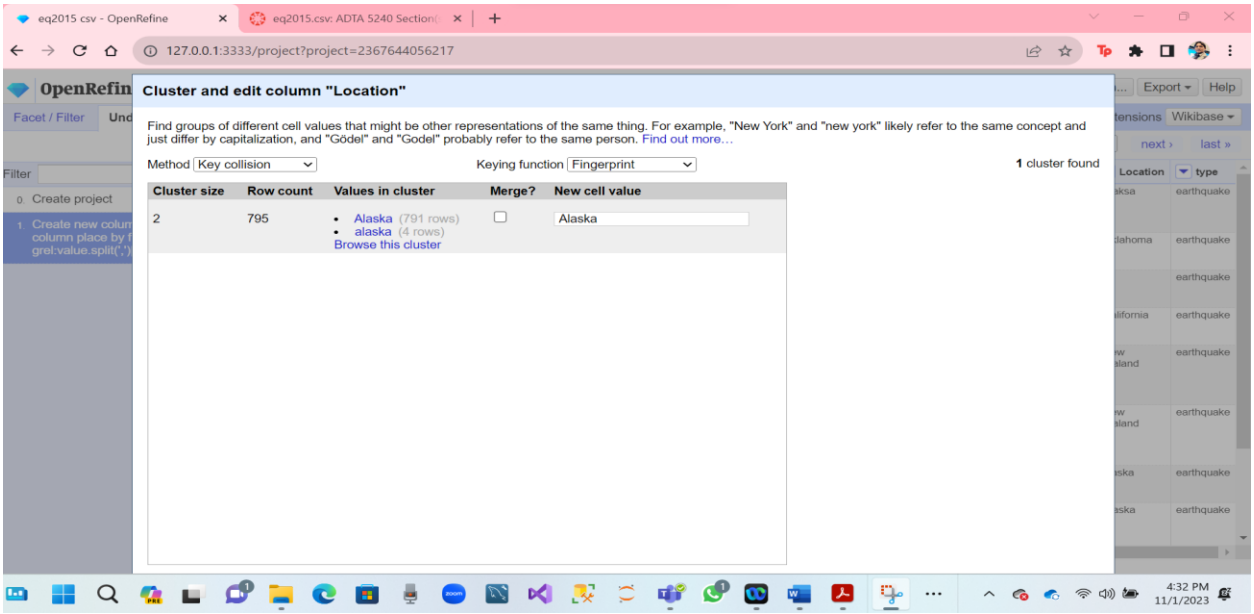
3. Exploring/Cleansing/Wrangling Data Set by Clustering

a. Using Earthquake 2015 project.

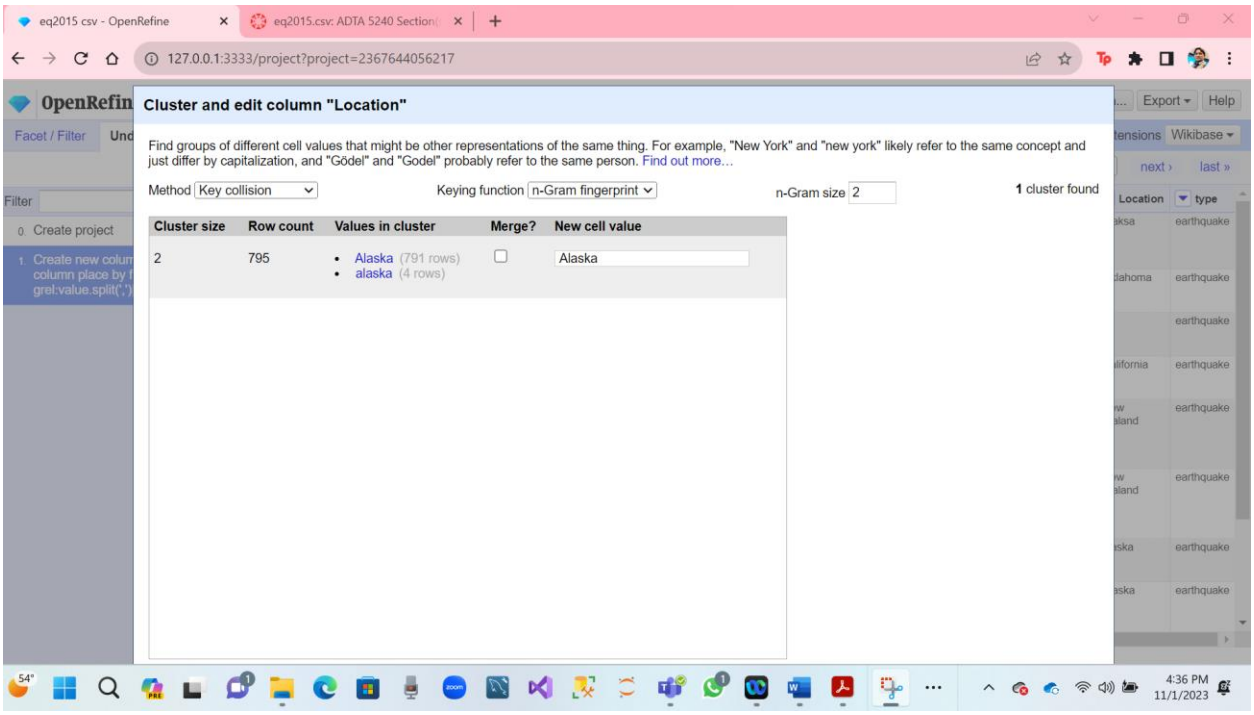
- I clicked on the drop-down menu arrow of location.
- Then I clicked on Edit cells and cluster and edit…
- After clicking on that I was taken to the below page.

b. Different ways to cluster

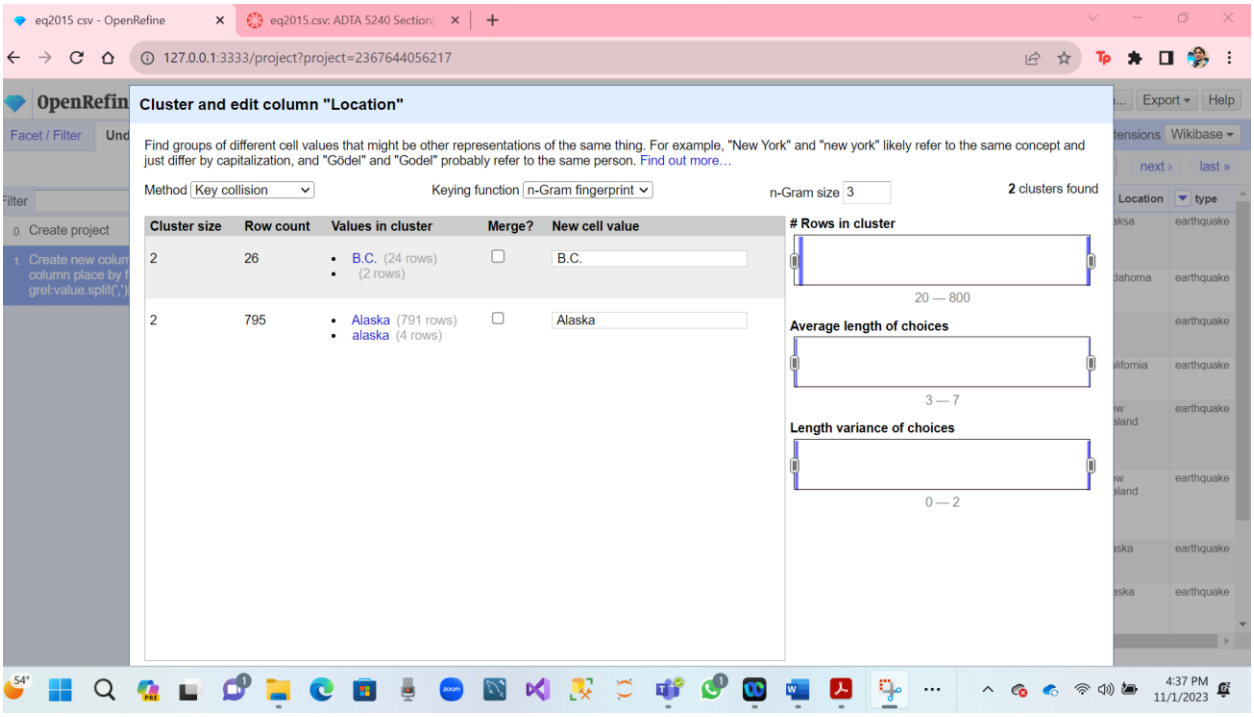
- OpenRefine: Clustering with the key collision method – fingerprint.



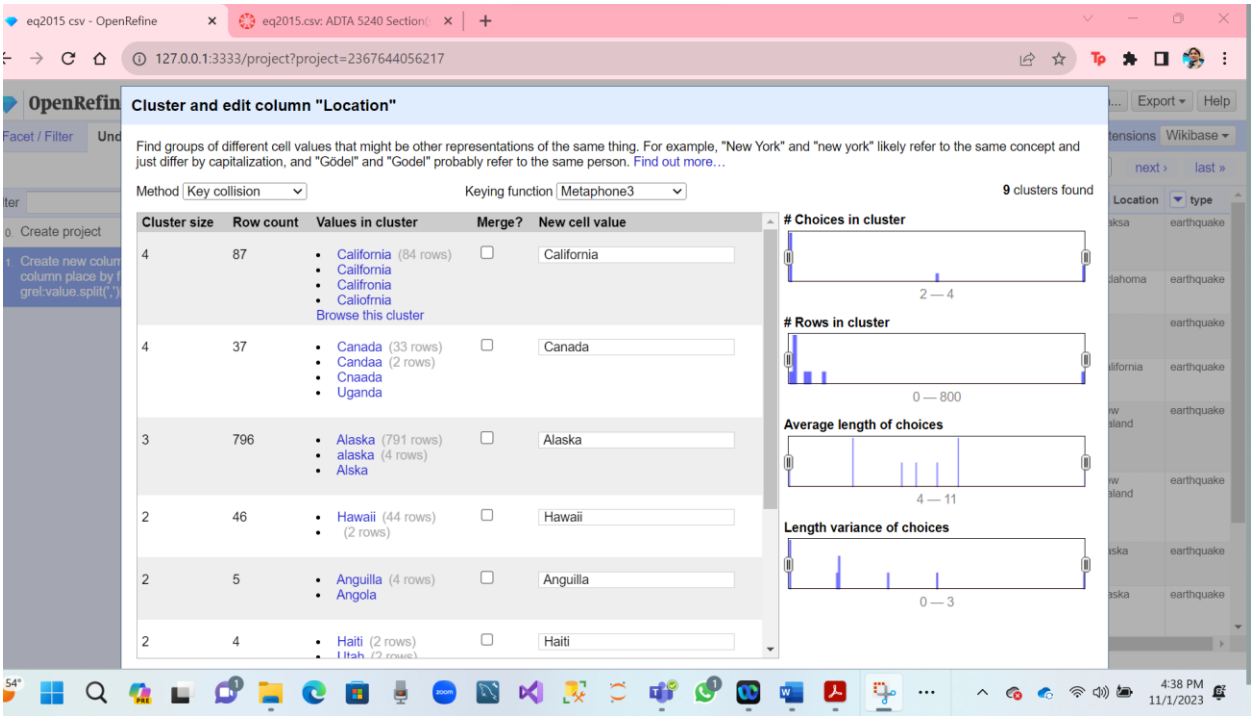
- OpenRefine: Clustering with the key collision method – n-Gram fingerprint with 2 for size.



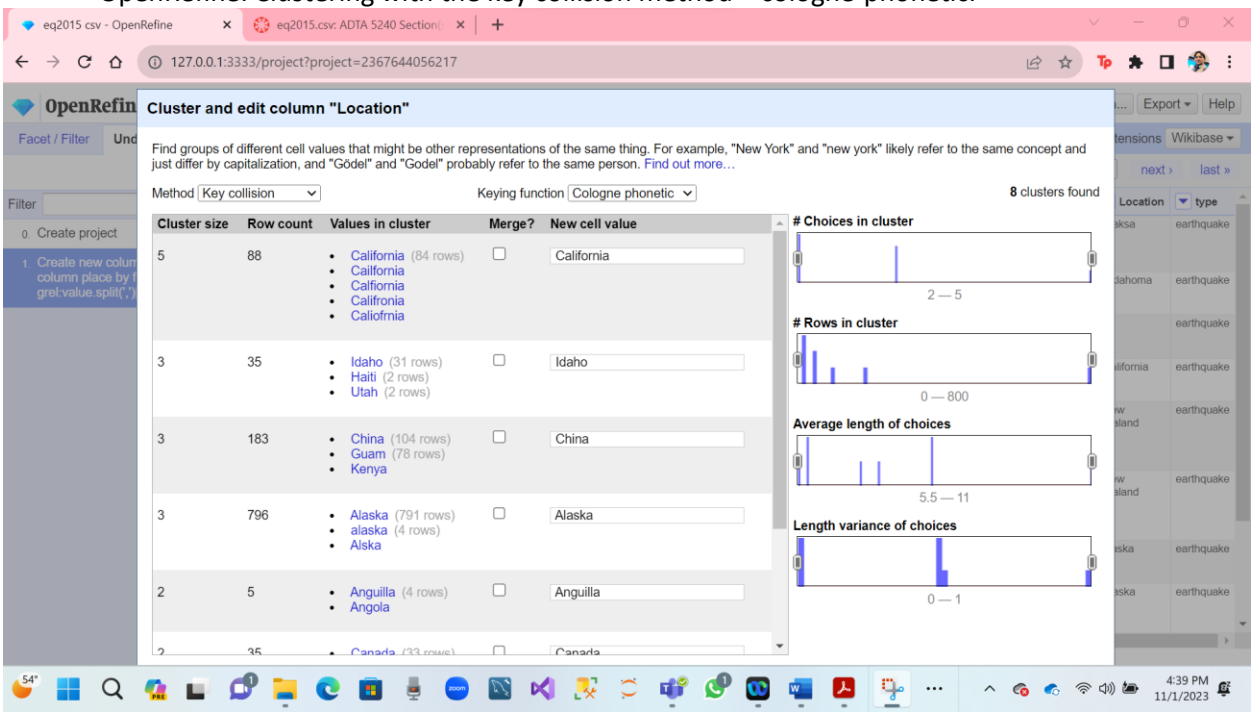
- OpenRefine: Clustering with the key collision method – n-Gram fingerprint with 3 for size



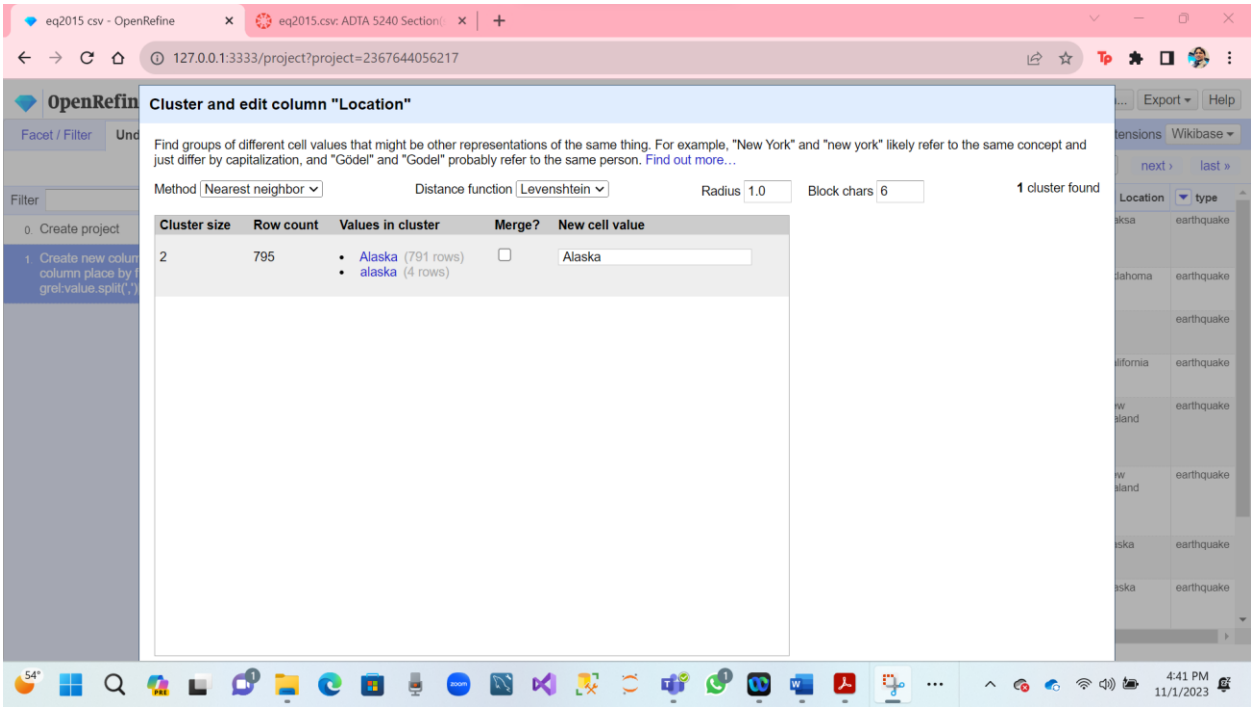
- OpenRefine: Clustering with the key collision method – metaphone3.



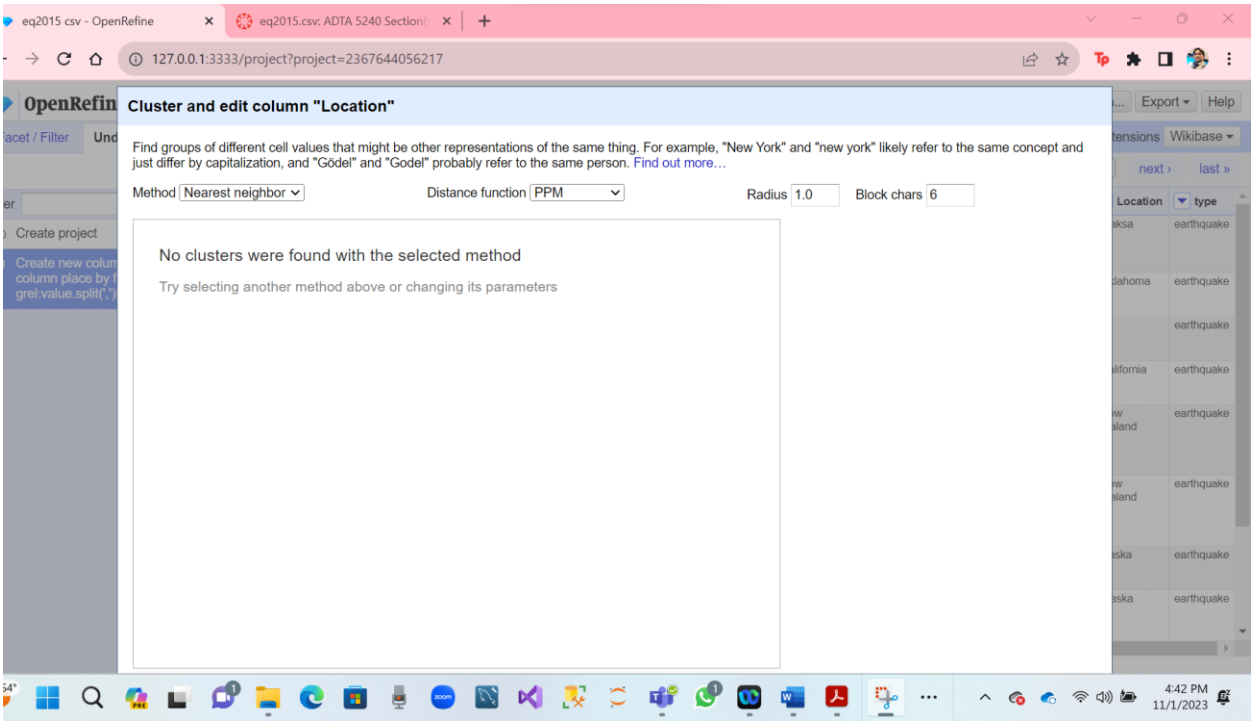
- OpenRefine: Clustering with the key collision method – cologne-phonetic.



- OpenRefine: Clustering with the nearest neighbour method: Levenshtein with radius of 1.0 and block chars of 6.



- OpenRefine: Clustering with the nearest neighbor method: Levenshtein with a radius of 2.0 and block chars of 6.
- OpenRefine: Clustering with the nearest neighbor method: Levenshtein with a radius of 3.0 and block chars of 6.
- OpenRefine: Clustering with the nearest neighbor method: PPM with a radius of 1.0 and block chars of 6.

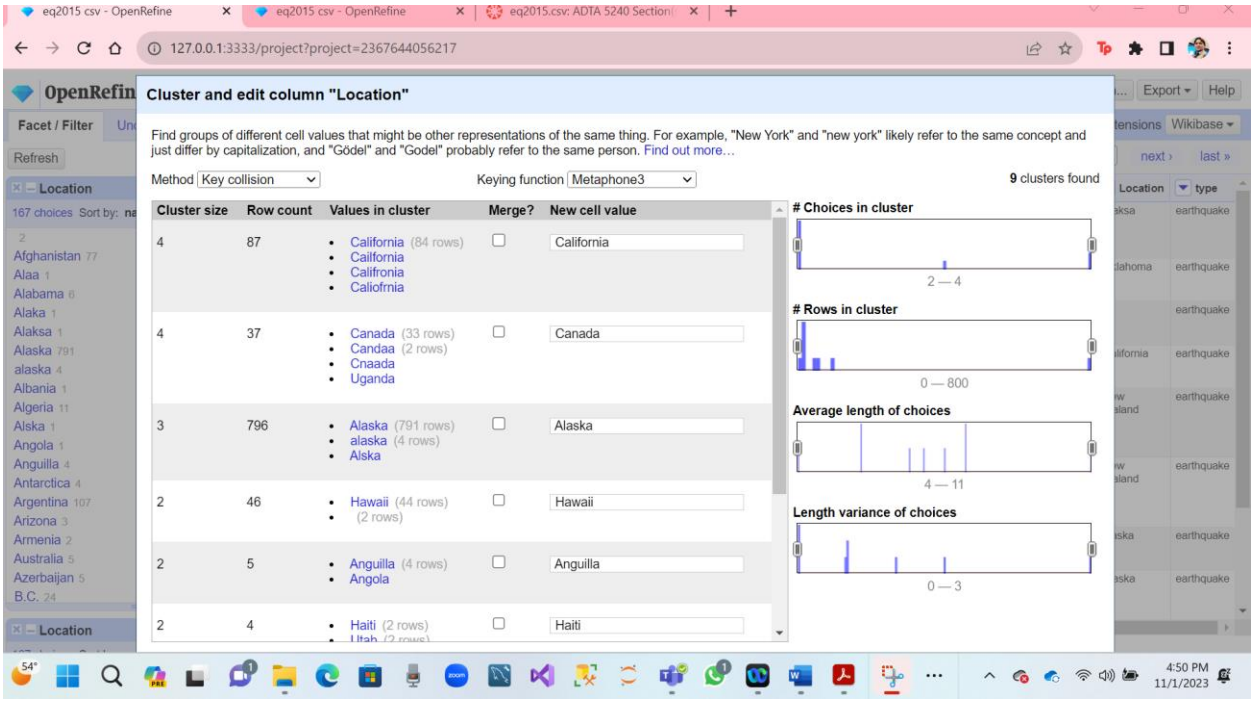


4. Cleansing/Wrangling data: Editing Cells

- Here we need to fix these types and errors using the clustering feature of OpenRefine.

Editing cells using the key collision method and metaphone3 keying function:

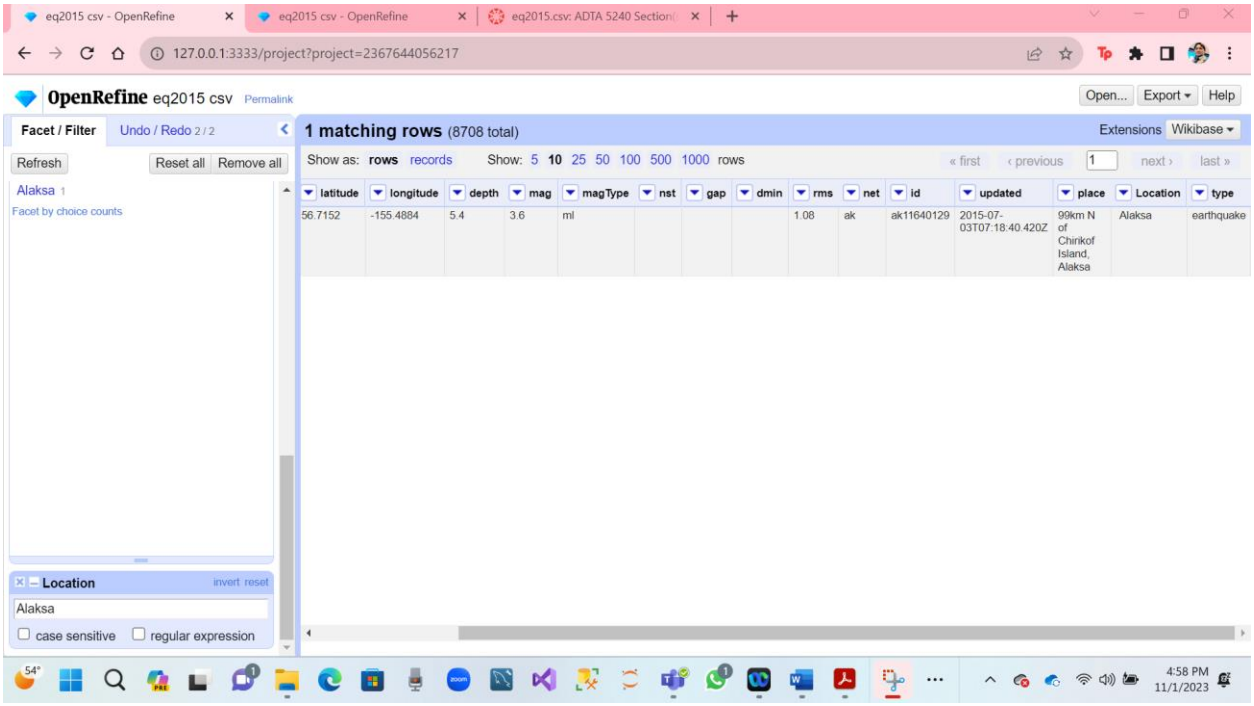
- Here I clicked on check boxes for the clusters California and Alaska.
- Then I clicked on merge.
- Here results show California and Alaska are gone.



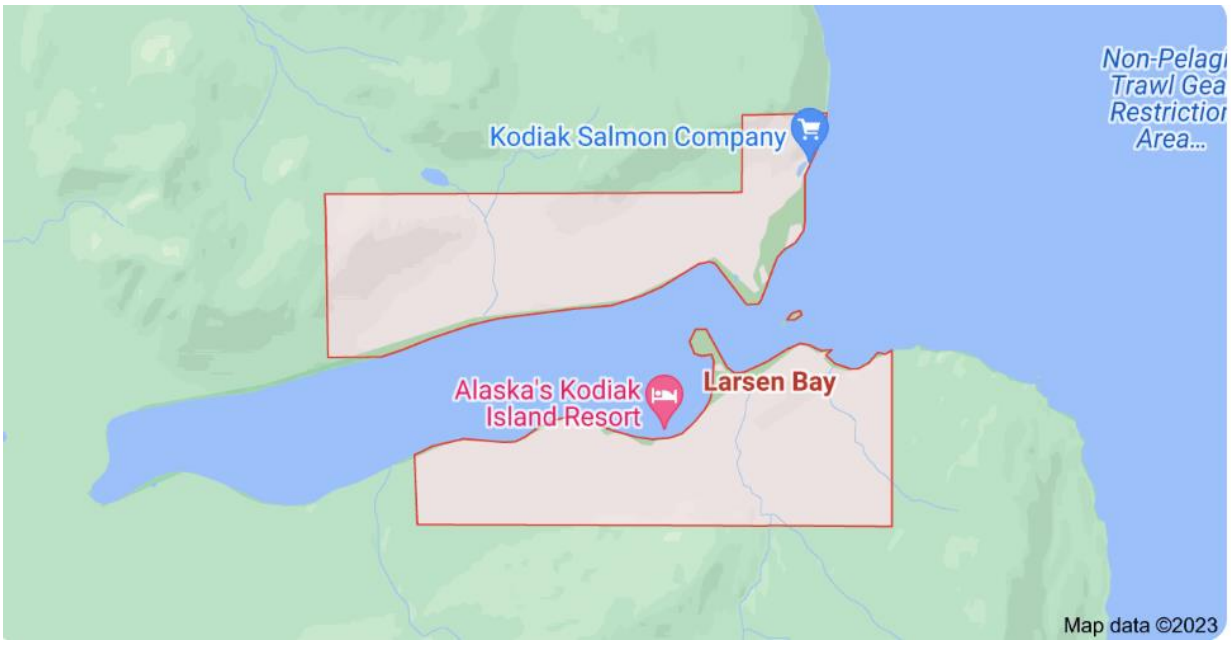
- Here the results show “Alaska” was capitalized as “A” now and “California” as “C”.

5. Cleansing/Wrangling Data: Manually editing cell:

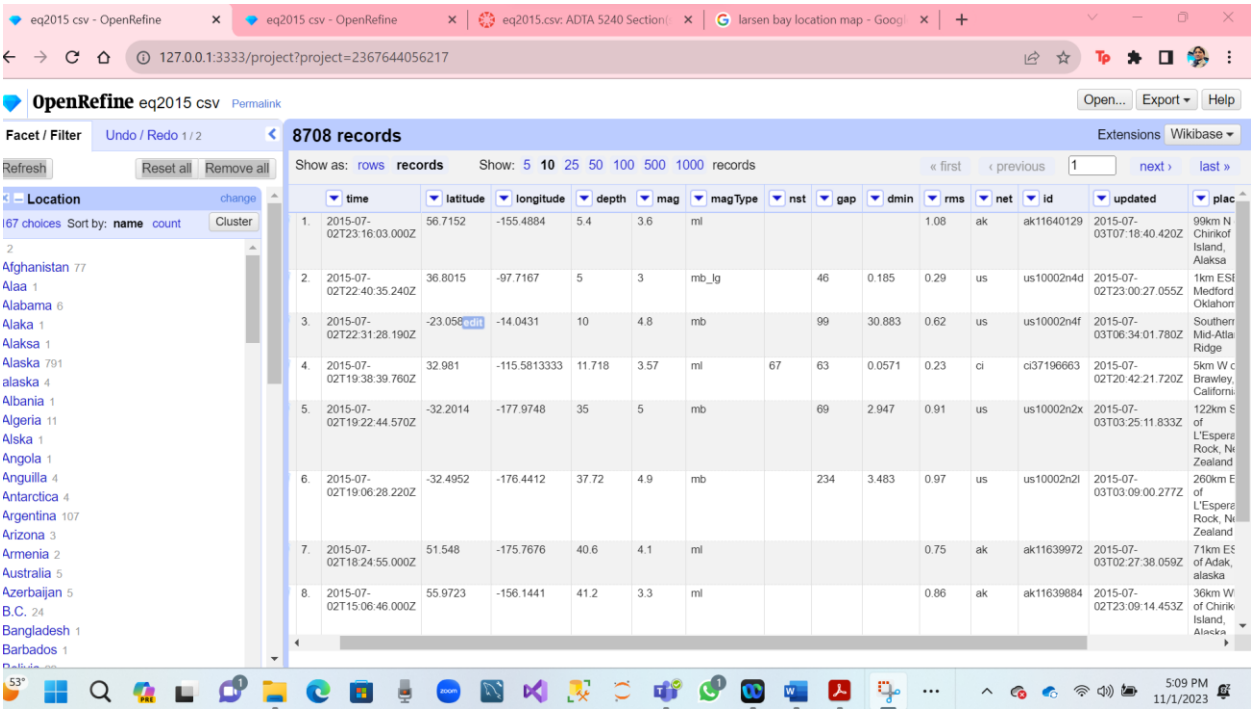
- Sometimes we need to edit the cells manually to correct the type of errors.
- Here we use “edit” feature of a cell to edit the text of the value.
- Then I clicked on undo and redo selected mass edit 8803 cells in column location.
- Then I clicked on Alaka.
- We will look at it now because we want to confirm that as Larson Bay is Alaska and not something else.
- I right clicked on Larson Bay in place column and clicked on google search.



- After clicking on that it took me to the below page.



- Then I came back to undo and redo in Open Refine page.
- Here I have undo all the functions previously I did.
- Hence, we have cleansed the data.
- It is OpenRefine with a couple of different data sets.



- Finally, Successfully Data ore with OpenRefine.