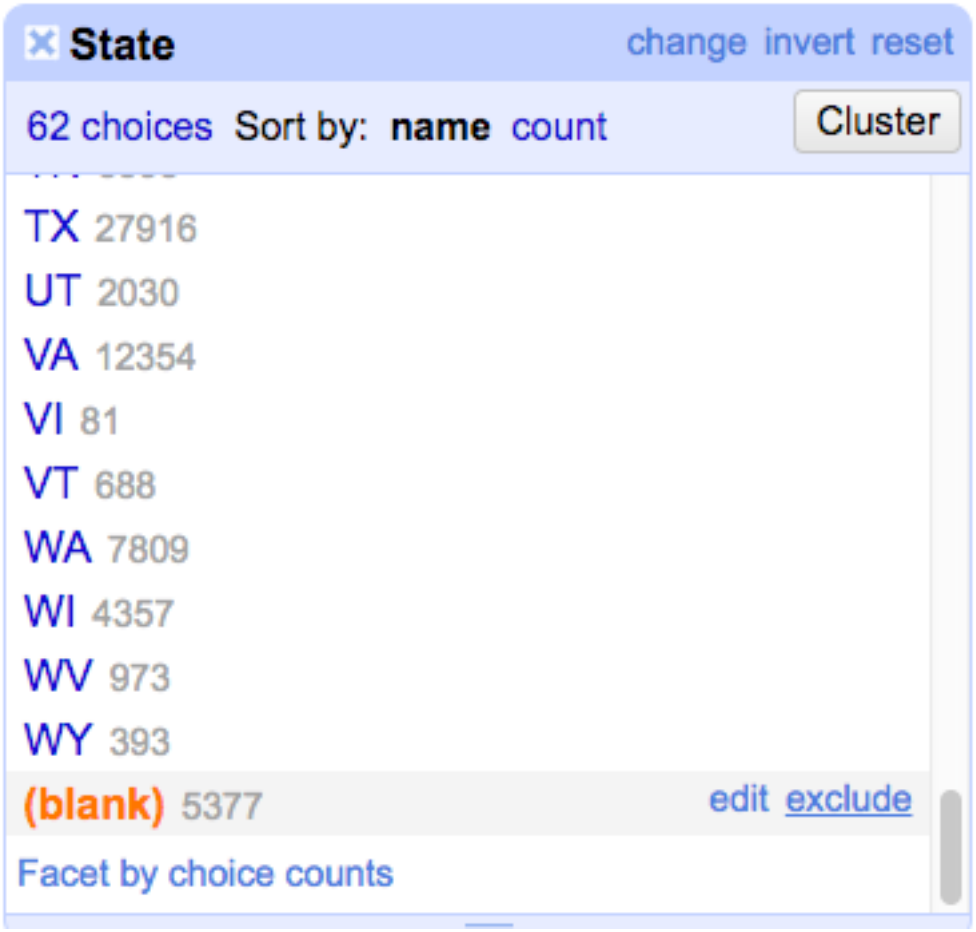
Lab 8- W205 Section 3 Spring 2017
Sanjay Dorairaj

Step1: Wrangling the Customer Complaints Data

Submission 1

How many rows are missing a value in the "State" column? Explain how you came up with the number.

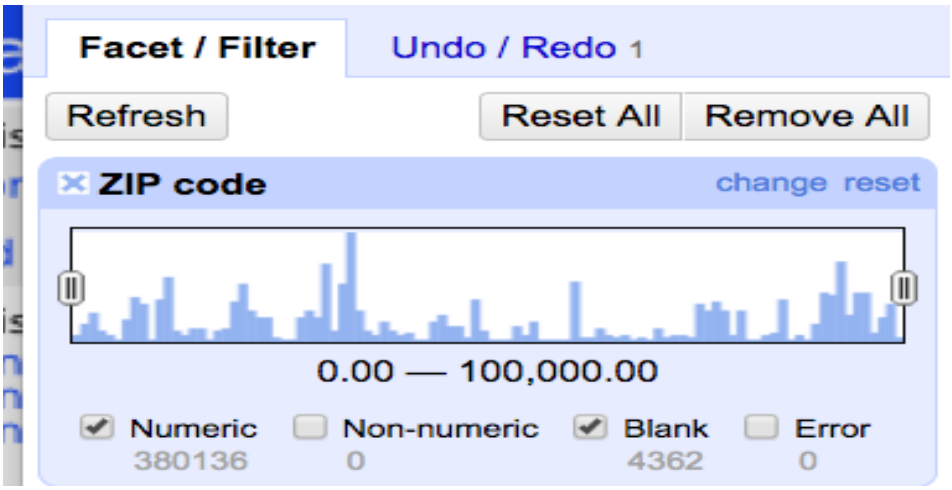
Number of rows missing a value in the State column are 5377



SUBMISSION 2:

How many rows with missing ZIP codes do you have?

There are 4362 rows with missing zip codes

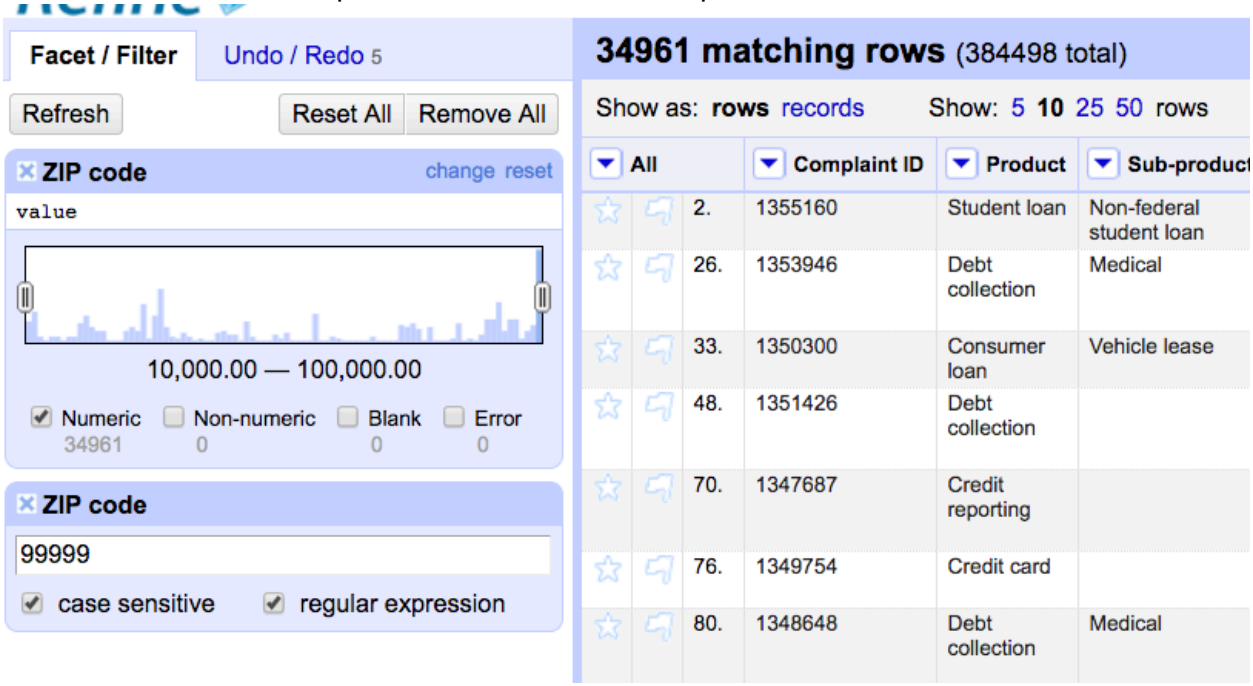


After applying fill-down, the blank cells assumed the zip code of the record immediately before that cell. The problem here is that the zip code may be wrong. One alternative approach may be to sort the column by zip code and then proceed with a fill-down. While this approach may still be prone to errors, it will have fewer errors than applying fill-down as-is.

SUBMISSION 3:

*If you consider all ZIP codes less than 99999 to be valid, how many valid and invalid ZIP codes do you have,
respectively?

There are 34961 invalid zip codes and 349537 valid zipcodes

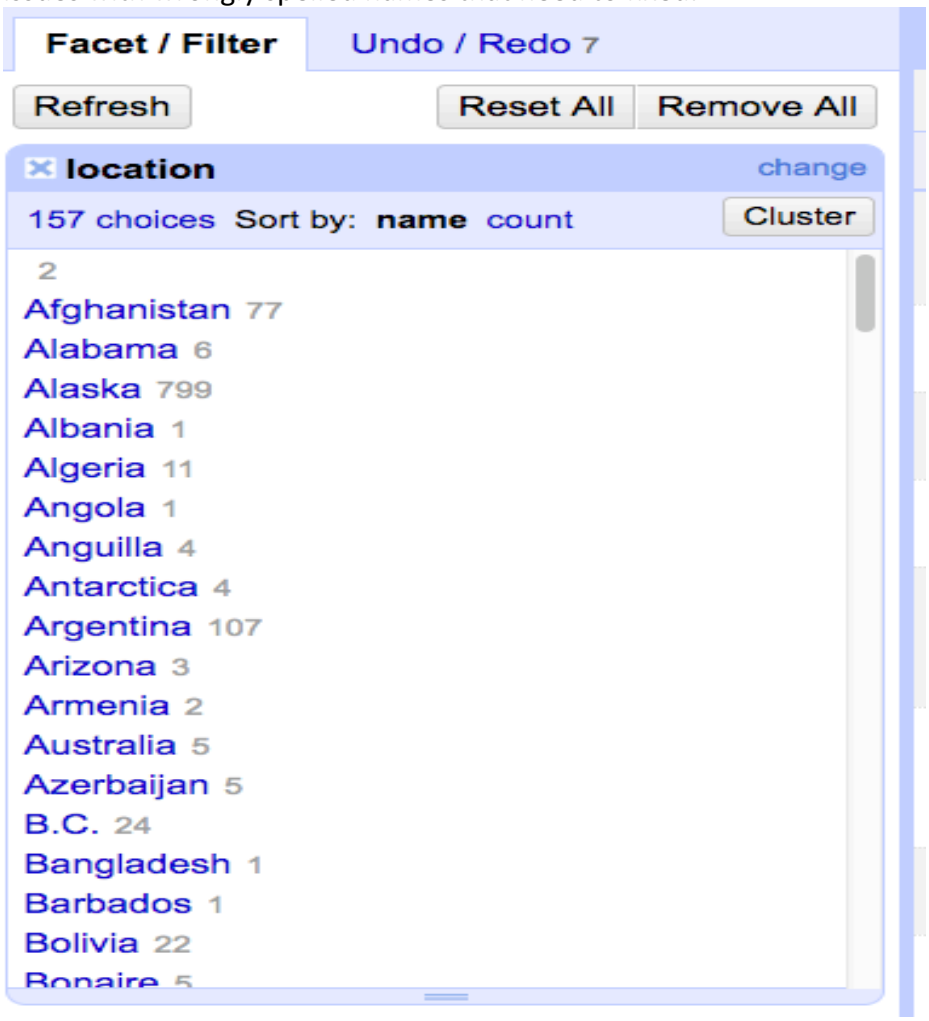


Step 2: Cleaning up eq2015 data

Nst – Number of seismic locations used to determine the earthquake location. Ignoring a row with missing nst values would cause us to lose all the other useful information gathered about earthquakes. One way of filling missing values may be to simply average existing values and use the average as a proxy for the missing values. One suggestion would be to simply replace missing values with the average value for nst.

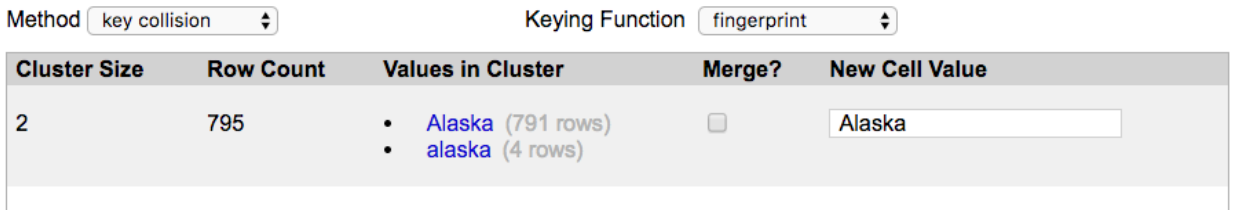
No clusters were found using the key collision method because of an earlier change during transformation in which all results were converted to TitleCase.

The new location is created per the instructions in the lab. Running a text facet shows several issues with wrongly spelled names that need to fixed.



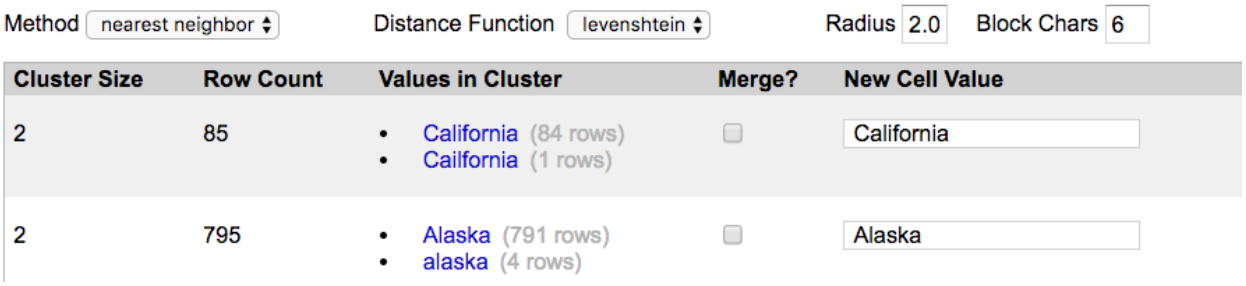
Using KNN and Levenshtein distance measure, with a radius of 2 and a blocksize of 4, several conflicts were found and result. The resulting records have 157 countries in them.

The Key Collision method using the fingerprint keying function yields the below cluster



Radius of 1 did not yield any results.

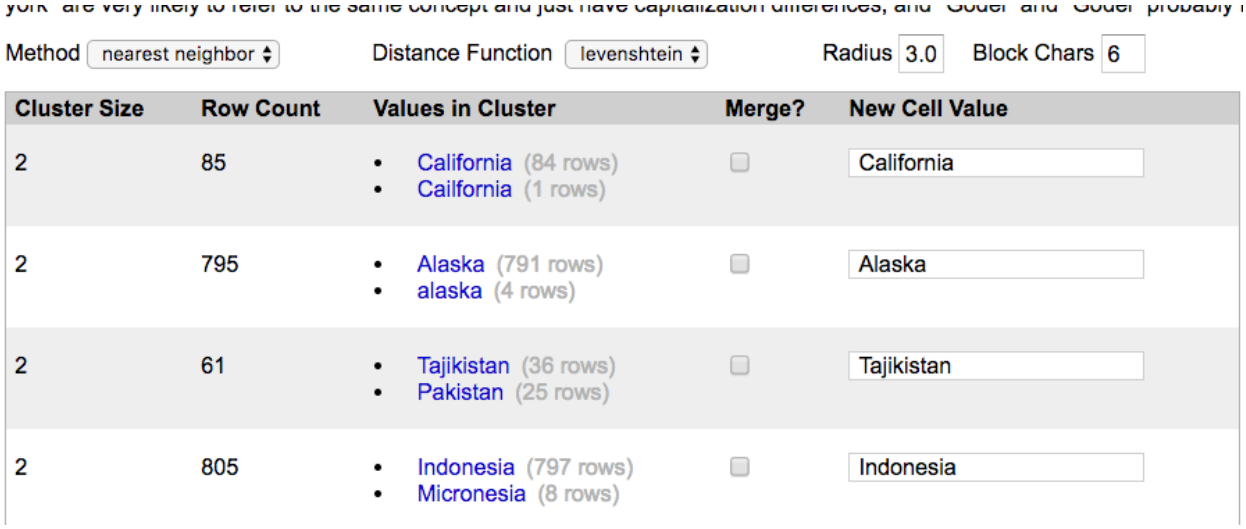
Using KNN and Levenshtein distance measure, with a radius of 2 and a blocksize of 6, results in the below clusters



SUBMISSION 4:

Change the radius to 3.0. What happens? Do you want to merge any of the resulting matches?

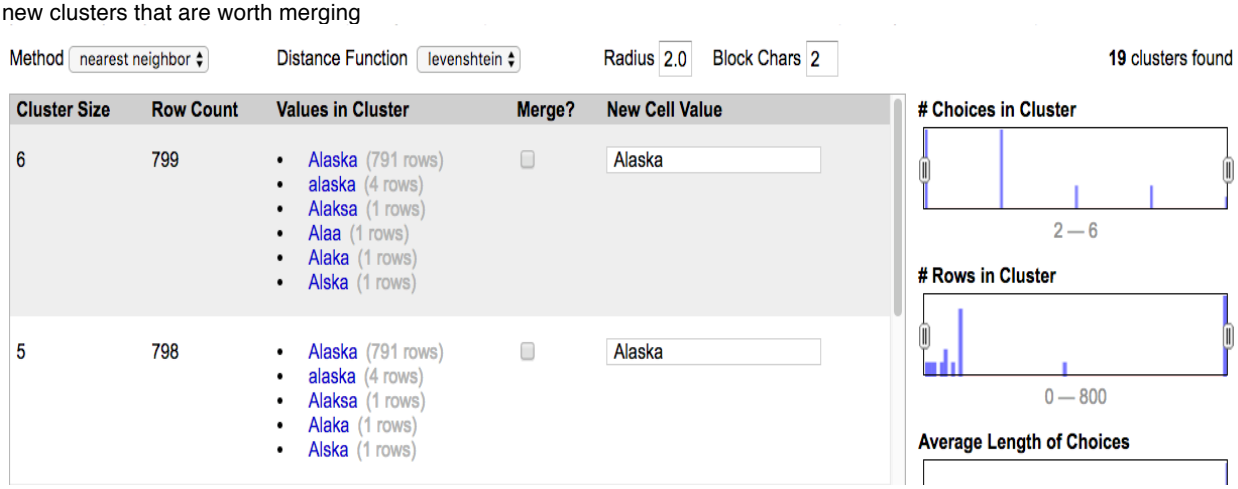
Changing the radius to 3 results in a few additional clusters, some of which do not include conflicting location names.



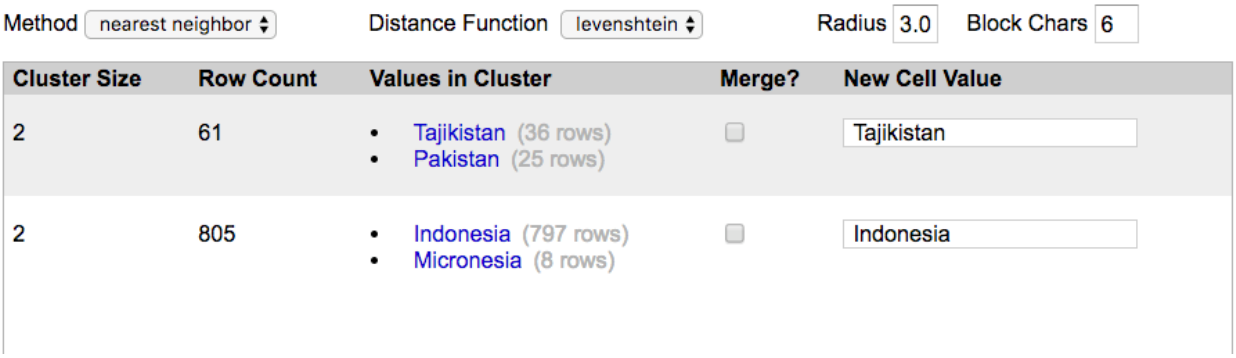
SUBMISSION 5:

Change the block size to 2. Give two examples of new clusters that may be worth merging.

Changing the radius to 2 reveals several new clusters (19 in all) several of which are worth merging. Shown below are two new clusters that are worth merging



After merging clusters, we are left with the below clusters



SUBMISSION 6:

Explain in words what happens when you cluster the "place" column, and why you think that happened. What additional functionality could OpenRefine provide to possibly deal with the situation?

Clustering of the “place” column appears to take an extraordinary amount of time. This may be because of the number of features that need to be modeled for cluster creation due to the size of the vocabulary. OpenRefine should provide an option to stop the execution of a task in case it takes a long time.

Step3 : Levenshtein Distance

SUBMISSION 7:

Submit a representation of the resulting matrix from the Levenshtein edit distance calculation. The resulting
value should be correct.

The below matrix gives the Levenshtein distance between gunbarrell and gumbarrel. The distance is 2. The computation was done in Excel using a formula. The same is attached. The value was verified using python (pdf attached)

		1	2	3	4	5	6	7	8	9	10
			G	U	M	B	A	R	R	E	L
1		0	1	2	3	4	5	6	7	8	9
2	G	1	0	1	2	3	4	5	6	7	8
3	U	2	1	0	1	2	3	4	5	6	7
4	N	3	2	1	1	2	3	4	5	6	7
5	B	4	3	2	2	1	2	3	4	5	6
6	A	5	4	3	3	2	1	2	3	4	5
7	R	6	5	4	4	3	2	1	2	3	4
8	R	7	6	5	5	4	3	2	1	2	3
9	E	8	7	6	6	5	4	3	2	1	2
10	L	9	8	7	7	6	5	4	3	2	1
11	L	10	9	8	8	7	6	5	4	3	**2**