SUBMISSION 1: How many rows are missing a value in the "State" column? Explain how you came up with the number.

There are 5,377 rows missing a value in the "State" column, because there are 5,377 blanks in the column "State" according to Open-Refine.

SUBMISSION 2: How many rows with missing ZIP codes do you have?

There are 4,362 blank zip codes or 4,362 rows with missing Zip codes when the "Numeric" is unchecked (or when only "Blank" is checked).

SUBMISSION 3: If you consider all ZIP codes less than 99999 valid ZIP codes, how many valid and invalid ZIP codes do you have, respectively?

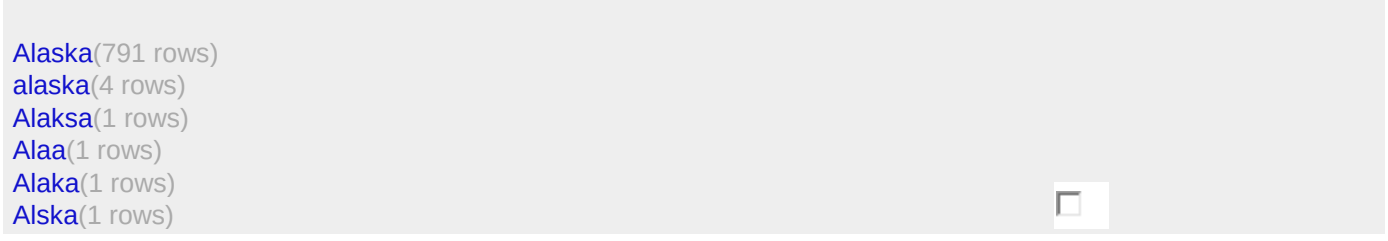
If all Zip codes less than 99999 are considered valid, then we have 380,119 valid zip codes and 4379 invalid zip codes, based on Open-Refine. (I added a column by "if(value < 99999, 1, 0)", then used a text facet.

SUBMISSION 5: Change the radius to 3.0. What happens? Do you want to merge any of the resulting matches?

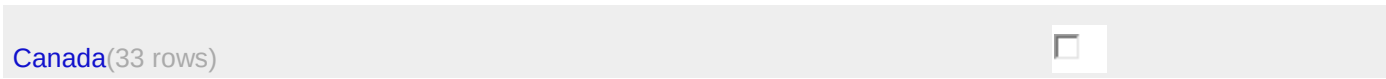
When radius is changed to 3.0, there are two new scenarios found by OpenRefine: "Tajikistan" vs. "Pakistan" and "Indonesia" vs. "Micronesia". Not like the original (when radius is 2.0) "Alaska" vs "alaska" and "California" vs. "Cailfornia"), I would not want to merge these two new 'matches' because Both are legitimate location names on themselves.

SUBMISSION 6: Change the block size to 2. Give two examples of new clusters that may be worthwhile merging.

When block size is changed to 2 (and the radius back to 2 also), I found more clusters worthwhile merging:



And



Candaa(2 rows)
Cnaada(1 rows)

And

California(84 rows)
Calfiornia(1 rows)
Cailfornia(1 rows)

SUBMISSION 7: 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?

Before the clustering, the text facet for "place" already yielding too many values to display. When "Cluster and Edit" is hit, there was no 'key collision' (of course), but when 'nearest neighbor' is chosen, the 'clock' just kept on turning and never came back (well more than 3 min). It is due to the involved calculation complexity (please note that it was m^2 pairwise, and there are O(n!) pairs!)

Perhaps OpenRefine can combine some default 'split' (divide and conquer) operation before pairing and applying Levenshtein. (That is apply some default subclustering (or even networked subclustering) first.)

SUBMISSION 8: Submit a representation of the resulting matrix from the Leveshtein edit distance calculation. The resulting value should be correct.

The distance is 3.

		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	E	7	6	5	5	4	3	2	2	2	3
9	L	8	7	6	6	5	4	3	3	3	2
10	L	9	8	7	7	6	5	4	4	4	3