Haroon Choudery
W205 Lab 8

SUBMISSION 1: How many rows are missing a value in the "State" column? Explain how you came up with the number.
There are 5377 rows that are missing a value in the State column. I came to this number by looking at the (blank) choice under the State facet.

SUBMISSION 2: How many rows with missing ZIP codes do you have?
There are 4362 missing values in the ZIP code facet. If you look at the numeric facet of the ZIP code data, you can see that there are 4362 values that are Blank next to the Blank checkbox.

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?
There are 34961 values that are equal to 99999 and the remaining 349537 are less than 99999. Of those, 349537, there are 4362 blank values so if we change the blank values to 99999, there are a total of total 39323 values of 99999.

SUBMISSION 4: Change the radius to 3.0. What happens? Do you want to merge any of the resulting matches?
There are four proposed matches - two of which are proper corrections (i.e. California and Cailfornia) and two of which are not proper corrections (i.e. Tajikistan and Pakistan)

SUBMISSION 5: Change the block size to 2. Give two examples of new clusters that may be worthwhile merging.
1. Canada, Candaa, and Cnaada
2. Alaska, alaska, Alaksa, Alaa, Alaka, Alska

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?
When you cluster the "place" column, the clustered get grouped together with similar locations and similar directions from the city. However, this is an issue because we do not want the clustering to make the beginning portion of the "place" values identical if they are in the same cluster. These numerical values should be unique to each value. It would be helpful if OpenRefine provided a way cluster only a portion of the column values together (perhaps using Regular Expressions). In this case, that would help preserve the portion before the comma for each value, as that should be unique.

SUBMISSION 7: Submit a representation of the resulting matrix from the Leveshtein edit distance calculation. The resulting value should be correct.

		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**