Lab 8

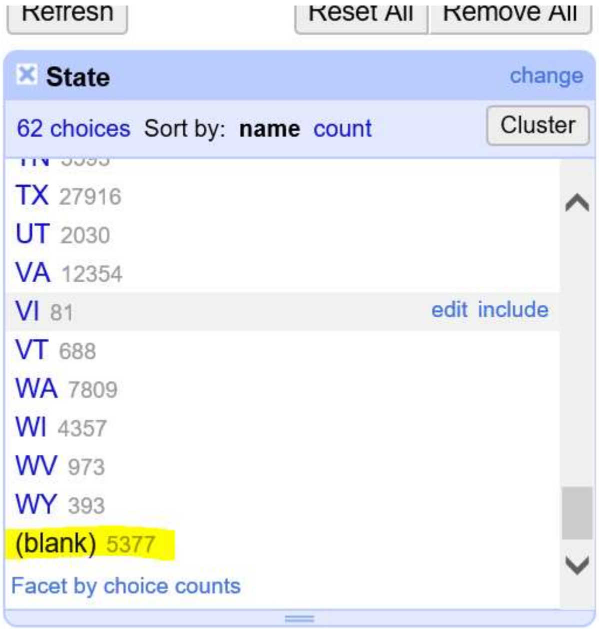
Cynthia Hu

Step 1. 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.

Answer:

There are 5377 rows missing a value in “State” column as there are 5377 rows with blank for “State” column from the State facet analysis.



State

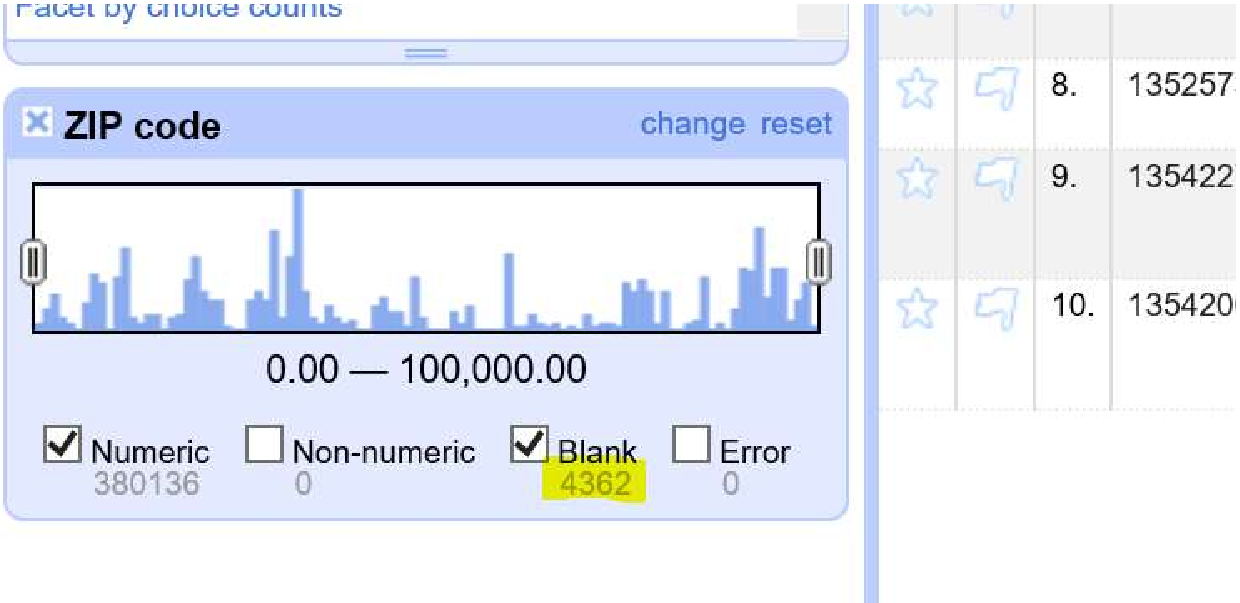
62 choices Sort by: **name** count

TX	27916
UT	2030
VA	12354
VI	81
VT	688
WA	7809
WI	4357
WV	973
WY	393
(blank)	5377

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

Answer:

I have 4362 rows with missing ZIP codes.



Facet by choice counts

ZIP code

0.00 — 100,000.00

☑ Numeric 380136
□ Non-numeric 0
☑ Blank 4362
□ Error 0

8. 135257
9. 135422
10. 135420

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?

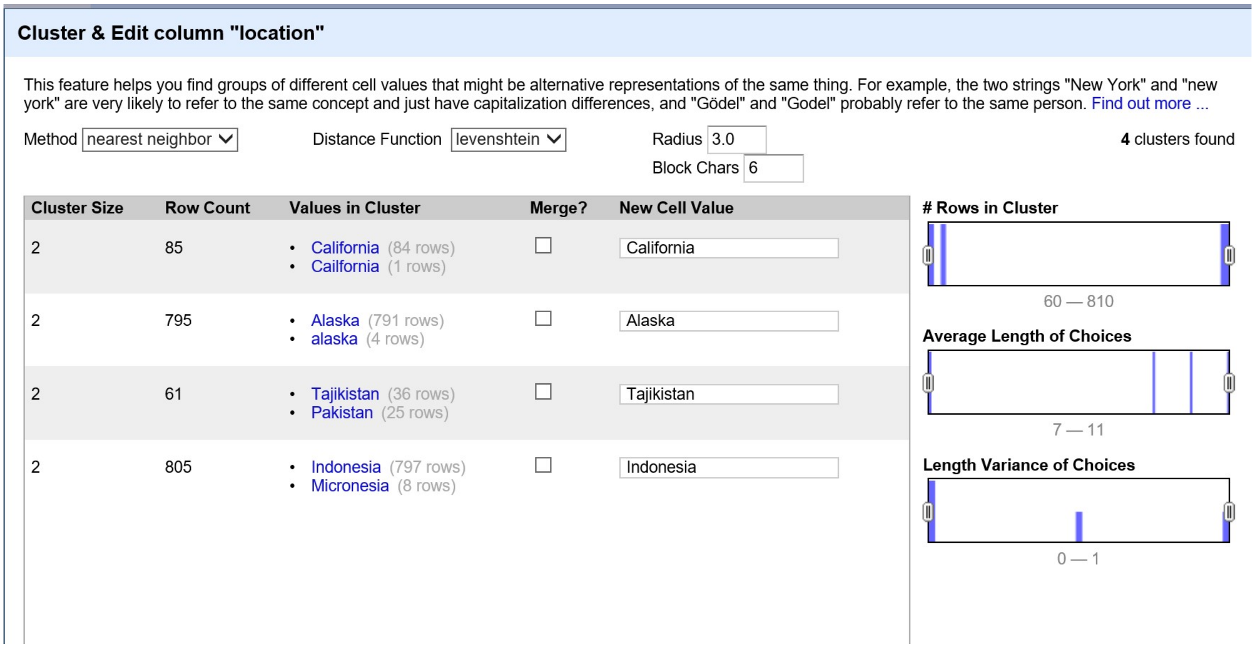
Answer:

Blank	4,362
Valid	345,175
Invalid	34,961
Total	**384,498**

Step 2. Cleaning Up eq2015 Data

SUBMISSION 4:
Change the radius to 3.0. What happens? Do you want to merge any of the resulting matches?

Answer:
When change the radius to 3.0, more results returned. However, I don't want to merge the additional results, like "Tajikistan" or "Indonesia".



Cluster & Edit column "location"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

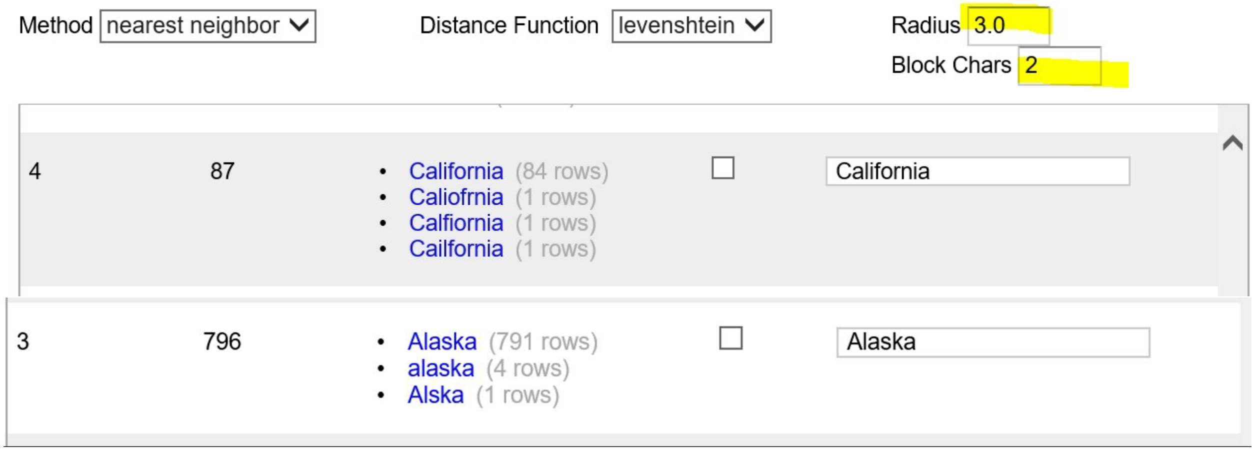
Method: nearest neighbor   Distance Function: levenshtein   Radius: 3.0   Block Chars: 6

4 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	85	• California (84 rows) • Cailfornia (1 rows)	□	California
2	795	• Alaska (791 rows) • alaska (4 rows)	□	Alaska
2	61	• Tajikistan (36 rows) • Pakistan (25 rows)	□	Tajikistan
2	805	• Indonesia (797 rows) • Micronesia (8 rows)	□	Indonesia

SUBMISSION 5:
Change the block size to 2. Give two examples of ***new clusters*** *that may be worth merging.*

Answer:
Below two new clusters may be worth merging.



Method: nearest neighbor   Distance Function: levenshtein   Radius: 3.0   Block Chars: 2

4	87	• California (84 rows) • Caliofrnia (1 rows) • Calfiornia (1 rows) • Cailfornia (1 rows)	□	California
3	796	• Alaska (791 rows) • alaska (4 rows) • Alska (1 rows)	□	Alaska

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?
Hint: you may want to cancel the run.

Answer:
When cluster the “place” column, there are no clusters found with the ‘key collision’ method. If use ‘nearest neighbor’ method, it’s keep running for a long time and no results returned. I think it’s because there are too many unique values in ‘place’ column and it takes much longer time to run the algorithm. OpenRefine may provide useful message during the process or kill the process automatically.

Step 3. Levenshtein Distance

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

Answer:
The distance between "gumbarrel" and "gunbarell" 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**