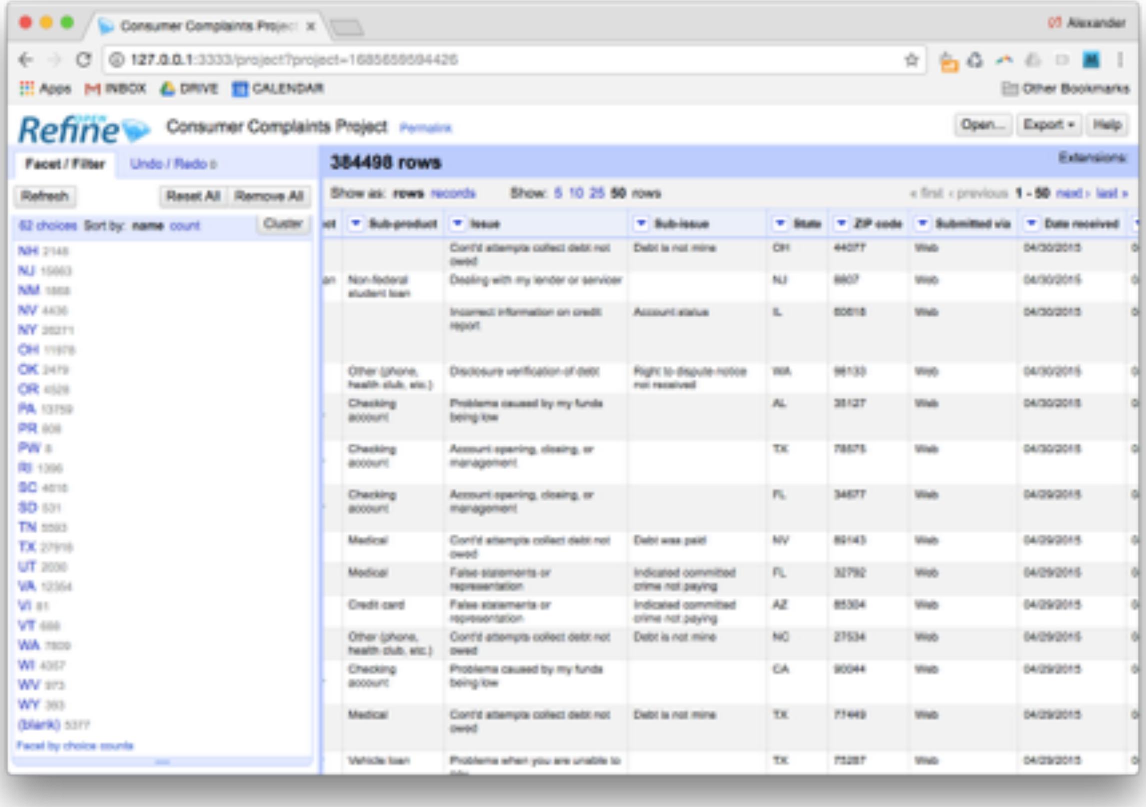
SUBMISSION 1:

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

It appears that **5377** rows return “Blank” for their state.

I came up by my answer by looking at the text facet of the state column. See left panel in below picture. I also sorted the data by the blank fields to double check they were blank.



Consumer Complaints Project Permalink

384498 rows

Show as: rows records Show: 5 10 25 50 rows

« first ‹ previous 1 - 50 next › last »

62 choices Sort by: name count

NH 2148
NJ 15663
NM 1808
NV 4436
NY 26271
OH 11978
OK 2479
OR 4528
PA 13759
PR 808
PW 8
RI 1396
SC 4016
SD 531
TN 5593
TX 27916
UT 2030
VA 12354
VI 81
VT 686
WA 7809
WI 4357
WV 973
WY 393
(blank) 5377

Facet by choice counts

ct	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received
		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015
an	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015
		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015
	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015
	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015
	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015
	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015
	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015
	Medical	False statements or representation	Indicated committed crime not paying	FL	32792	Web	04/29/2015
	Credit card	False statements or representation	Indicated committed crime not paying	AZ	85304	Web	04/29/2015
	Other (phone, health club, etc.)	Cont'd attempts collect debt not owed	Debt is not mine	NC	27534	Web	04/29/2015
	Checking account	Problems caused by my funds being low		CA	90044	Web	04/29/2015
	Medical	Cont'd attempts collect debt not owed	Debt is not mine	TX	77449	Web	04/29/2015
	Vehicle loan	Problems when you are unable to		TX	75287	Web	04/29/2015

SUBMISSION 2:

How many rows with missing ZIP codes do you have?

It appears that **4362** rows are blank. This is *after* transforming the zip code text into number format.

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?

I don't completely understand this question but here's what I think it's saying:

```
if zip code == 99999:
        zip code is INVALID

if zip code < 99999:
        zip code is VALID

if zip code == "Blank"
        zip code is INVALID

Total INVALID = count(99999) + count(blank)
                        =  34961 + 4362
                        =  39323
```

So, total invalid is **39,323**. Total valid would be **345,175** (total rows *minus* total invalid or 384498 - 39323).

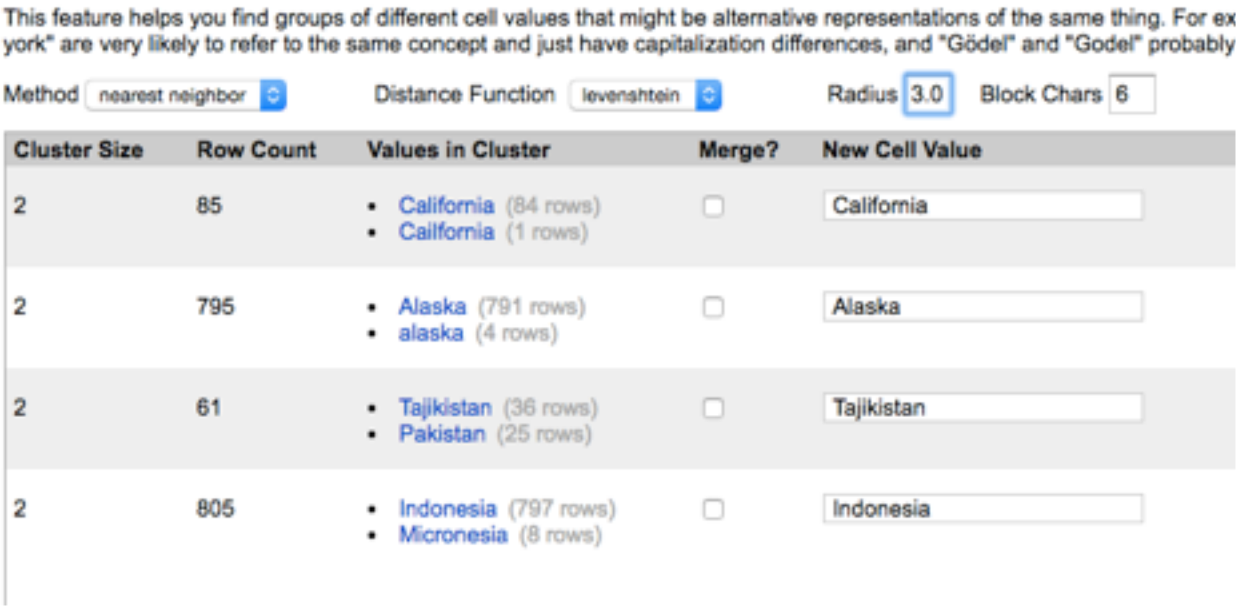
SUBMISSION 4:

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

When we changed the radius to 3.0, the cluster began to pair words that were similar but were actually distinct.

So, for example, Indonesia and Micronesia are NOT misspellings but rather two different regions with very similar spelling.

You don’t want to merge these particular values.



This feature helps you find groups of different cell values that might be alternative representations of the same thing. For ex
york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably

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

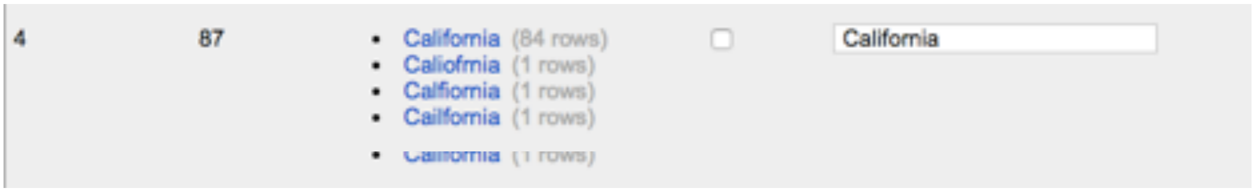
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	85	California (84 rows) Cailfornia (1 rows)	□	California
2	795	Alaska (791 rows) alaska (4 rows)	□	Alaska
2	61	Tajikistan (36 rows) Pakistan (25 rows)	□	Tajikistan
2	805	Indonesia (797 rows) Micronesia (8 rows)	□	Indonesia

SUBMISSION 5:

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

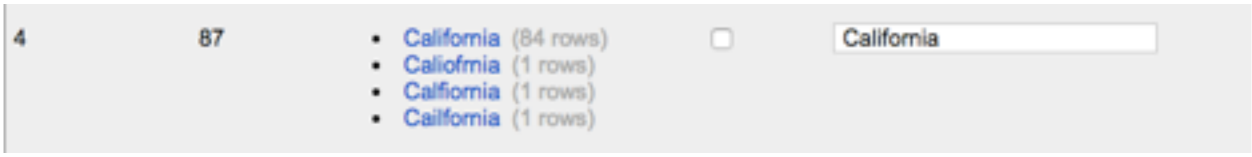
First: Two clear misspellings of California:

Cluster 1



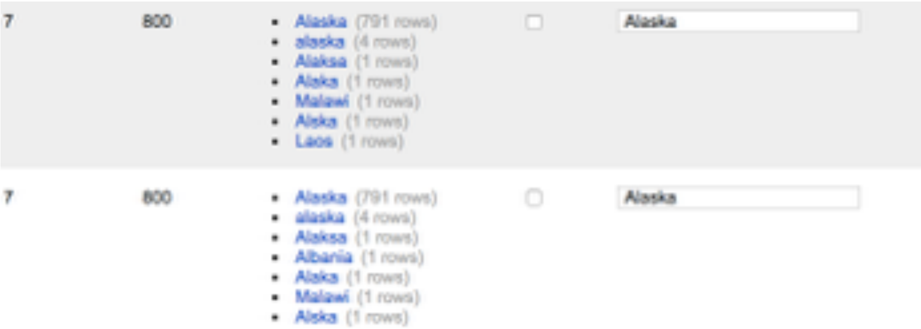
4	87	• California (84 rows) • Caliofmia (1 rows) • Calfiornia (1 rows) • Cailfornia (1 rows) • California (1 rows)	☐	California

Cluster 2



4	87	• California (84 rows) • Caliofmia (1 rows) • Calfiornia (1 rows) • Cailfornia (1 rows)	☐	California

Second: Two groupings that seem to be primarily Alaska. This would distort the data by folding in entries from 3 other regions. It was the “best worst” option presented in the group.



7	800	• Alaska (791 rows) • alaska (4 rows) • Alaksa (1 rows) • Alaka (1 rows) • Malawi (1 rows) • Alska (1 rows) • Laos (1 rows)	☐	Alaska
7	800	• Alaska (791 rows) • alaska (4 rows) • Alaksa (1 rows) • Albania (1 rows) • Alaka (1 rows) • Malawi (1 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?

The cluster seems to take a long time to process. Waited a few minutes before canceling.

This could have happened because:

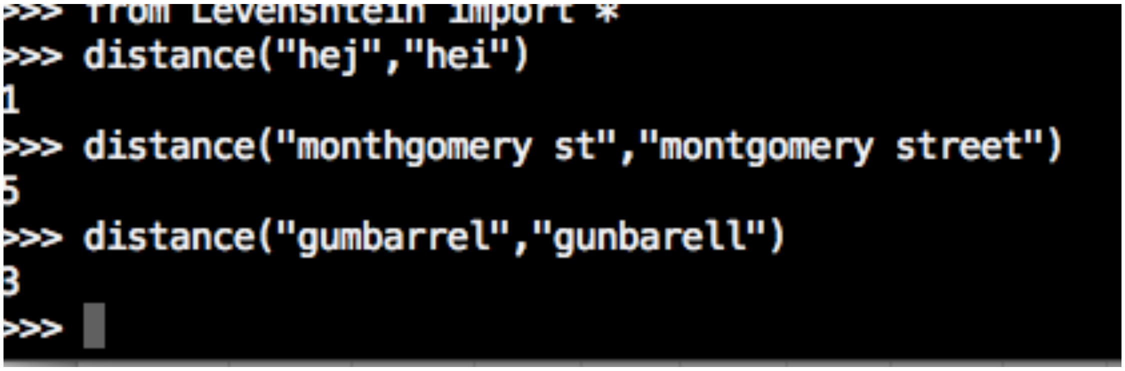
a. There were too many individuals terms to process
b. The terms were different data types
c. There was too much data to process
d. The clusters were so small and so populous that it was running for a long time before completing

OpenRefine could provide a few things:

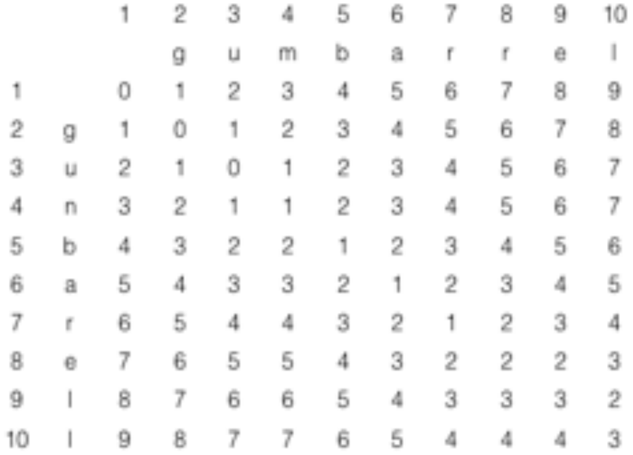
a. Split function that groups similar data types
b. A lighter weight visualization tool that just shows you all the unclustered data color coded by type
c. General rules about ideal data sets to cluster

SUBMISSION 7:

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



```
>>> from Levenshtein import *
>>> distance("hej","hei")
1
>>> distance("monthgomery st","montgomery street")
5
>>> distance("gumbarrel","gunbarell")
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