Lab 8

Cynthia Hu

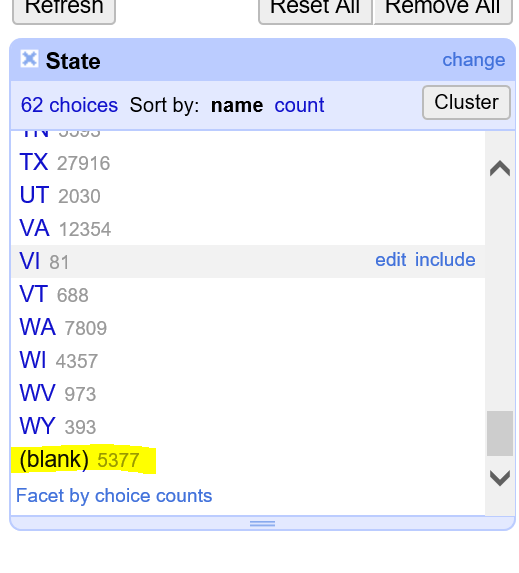
# Step 1. Wrangling the Customer Complaints Data

**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.

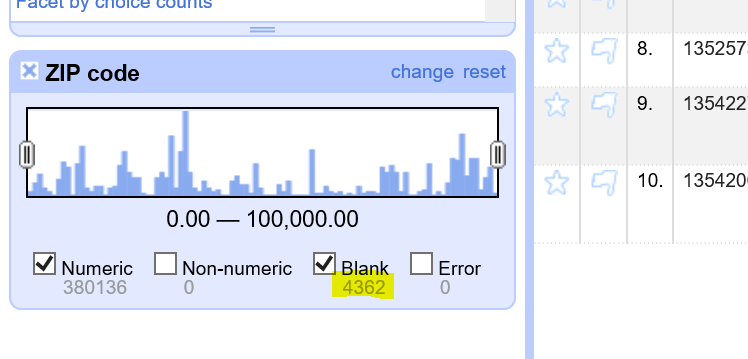


**SUBMISSION 2:**

*How many rows with missing ZIP codes do you have?*

**Answer:**

I have 4362 rows with missing ZIP codes.



**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** |

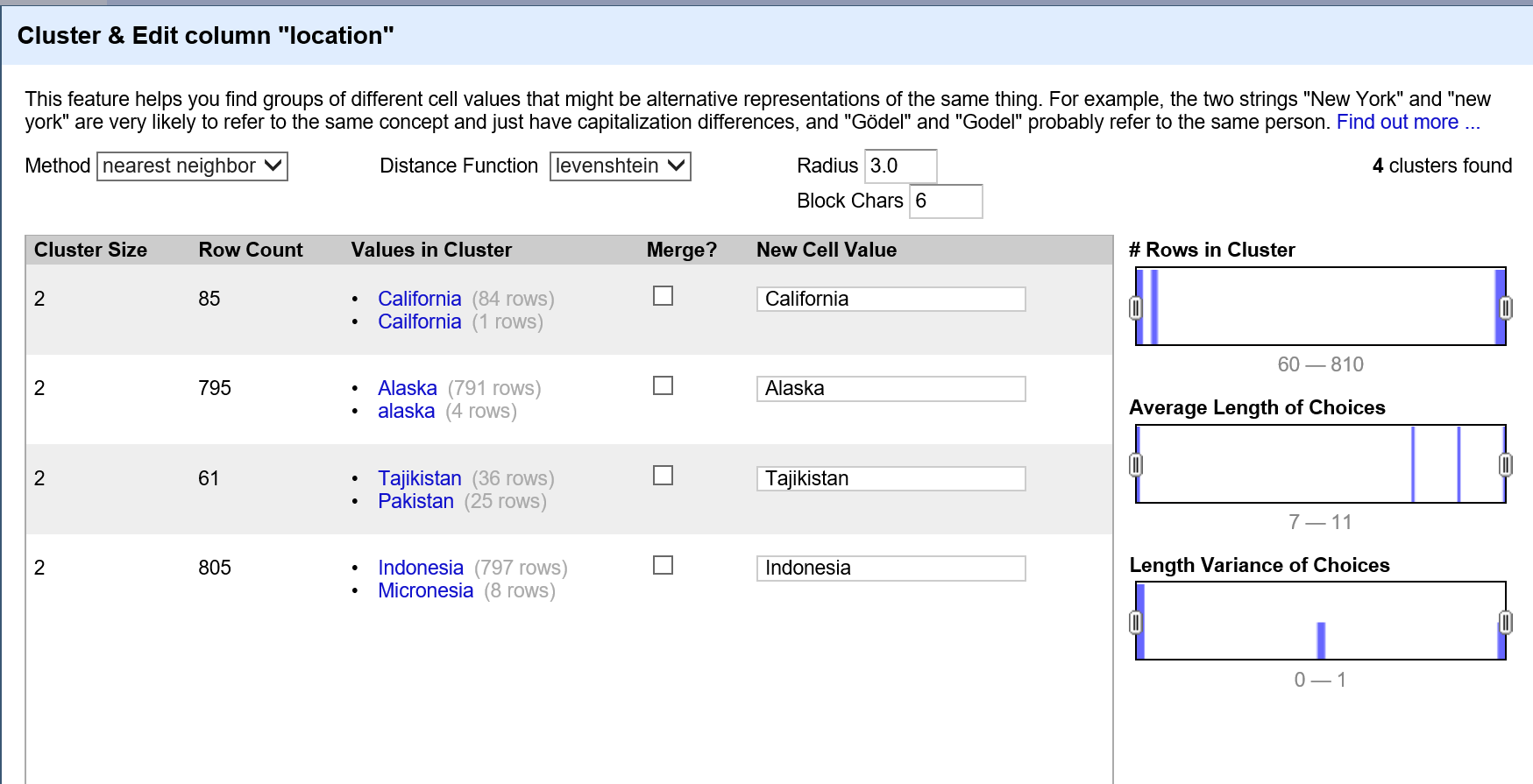
# Step 2. Cleaning Up eq2015 Data

**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”.

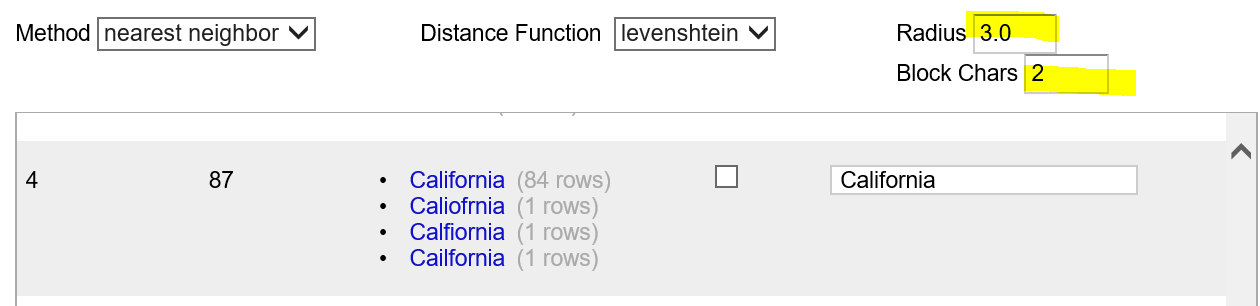


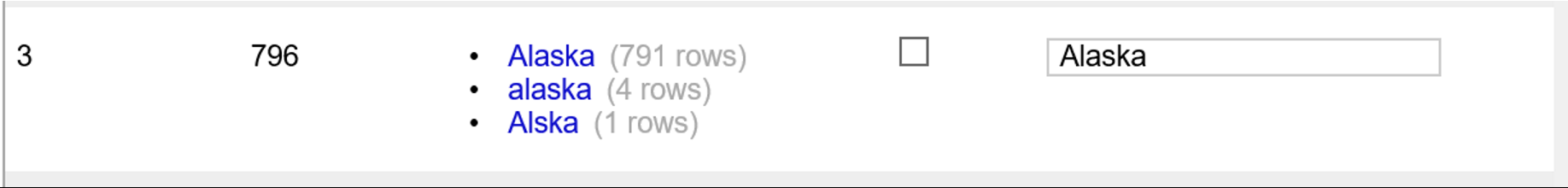
**SUBMISSION 5:**

*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.





**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** |