**Lab 10**

**Submission 1**

The reason it has 62 state values is because it includes 51 states, 10 other territories (i.e. when sort by count, the last from AE to AP stands for areas such as American Samoa (AS), Virgin Islands (VI), Guam (GU), Palau (PW), and US Armed Forces – Americas (AA), Europe (AE), and Pacific (AP)), and 1 blank.

See: <https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations>

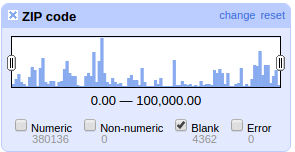
There are 5377 rows missing a value in the “State” column, this can be seen by the last entry in the summary – the count associated with value “(blank)” is 5377.



**Submission 2**

24748 unique zip codes seem too much and when eyeballing it, several zip codes (such as 1, 10, 100, 707, 7070) only have a single count. This doesn't make intuitive sense and does not follow US postcode standards.

It can be treated as numeric as the non-numeric value count is 0. There are in total 4362 rows with missing zip codes.



**Submission 3**

The problem with if(value.length() > 4, value, “99999”) is that the data in the new columns does not have the same type.

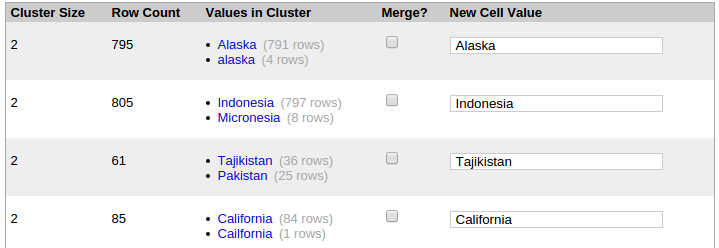


If consider all zip codes less than 99999 as valid, there are in total 345175 valid zip codes and 39323 invalid ones.



**Submission 5**

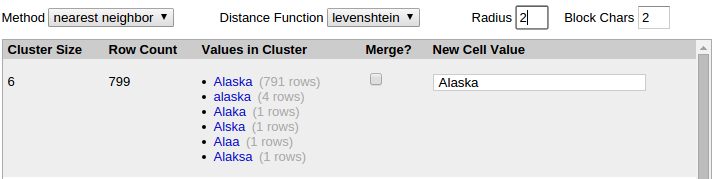
Since nst stands for “the total number of seismic stations which reported P- and S-arrival times”, a common strategy would be filling the blanks with 0s.

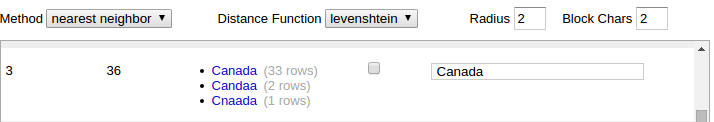
****

When changing the radius to 3.0, the clustering identified “Tajikistan” and “Pakistan”, and “Indonesia” and “Micronesia” as clusters, in addition to Alaska and California, which are valid at radius 1.0 and 2.0, and no, wouldn't want to merge the two new clusters that emerge with radius 3.0.

**Submission 6**

Below are screenshots of two examples of new clusters that I think are worth merging.

****

****

**Submission 7**

When using “nearest neighbour” on the “place” column, Google Refine takes a long time to compute – so long that one had to cancel it.

The reason this happened is due to the nature of the “nearest neighbour” algorithm – it does not require any training but prediction time is O(N^2), and since the “place” column contains much longer strings for each entry compared to the extracted value “location”, the computation time easily blows up.

One addition to Refine could be to pre-scan the data on which the user intends to perform clustering using “nearest neighbour” and estimate the time, or at least determine if it is suitable, and alert the user in case “nearest neighbour” is not the most suitable algorithm.

**Submission 8**

The Levenshtein distance is 3.