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

5377 rows are missing a value in the “State” column. I found this by looking at the count for the value “blanks”

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

4362 rows are missing ZIP codes

3) *If you consider all ZIP codes less than 99999 to be valid, how many valid and invalid ZIP codes do you have, respectively?*

349133 valid zip codes

35363 invalid zip codes

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

I’d only want to merge the California and Alaska clusters (not Tajikistan/Pakistan Indonesia/Micronesia)

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

Ex1 alaska, alaska, Alaska
Ex2 alaka, alska, alaksa, alaska, Alaska

6). Clustering this way is very inefficient. The process does not complete quickly (if ever!!!). The individual choice count is too high and suffers from long strings, which is likely a leading cause of this issue.

It would be helpful to first group by another column to narrow the search for clusters and run that process for each group simultaneously.

7). The answer is 3 (see next page for matrix)

		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**		

d[i,j] is to be set to the minimum of:

- d[i-1,j]+1
- d[i,j-1]+1
- d[i-1,j-1]+cost[i,j]