* Removing Duplicates

During the data clean-up process, we notice there are many records having duplicate names with different id\_person\_dim, address, states. But it is still very possible the same person moving around. For example:

2902 Adams Richard 2717 102 Bryan Ct. Apt. 101 Laurel MD US 10810

2458 Adams Richard 2304 Rt. #1, Box 439 Oakiland MD US 21550

In this case, the issue dates are two years apart. It's reasonable to assume that Adam could have moved from Laurel, Maryland, to Oakland, Maryland. So we work out the following rule to determine: if the person still live in the same state and the publication\_dates’ difference is within 2 years, then we consider him/her the same person.

For the actual workflow, we extract all duplicate name with id to a separate new file (901 records). Then searching through the original file by id, inspecting by the rule, and merging ids. Finally, the cleaned file is able to help us create visualizations without further confusion.