Exercise 1: Exploring Splitting the Data for Entity Resolution

February 1, 2021

Consider the `RLdata500` data set in the `RecordLinkage` package.

Suppose that this data set is too large to work with and we would like to create a sample size of 10 records that is representative of the 500 records in this data set. Let's investigate how we can do this!

1. Start by doing simple exploratory data analysis of the data set. What do you find?
2. What happens if you randomly sample 10 records from the original data set? Do this a few times and describe what happens? Is this representative of the original data set? Explain and be specific.
3. Propose something that works better than random sampling and explain why this works better.
4. Propose evaluation metrics, visualizations, etc to support any of your claims.

Write up what you do in class today with your group for homework 1 to help solidify your knowledge of entity resolution and push this to your class repository.

Hint 1: Think about the fact that you have unique identifiers and you know the maximum cluster size here.

Hint 2: Talk to myself/Olivier if you're getting stuck. This is a hard but very important problem!

Suggestion: Redo this for your homework 1 for El Salvador as something extra if you have time or want an extra challenge! Bonus points will be awarded for this if you do this extra task with your group!