Homework 3

STA 325: Understanding Entity Resolution Data

General instructions for homeworks: Your code must be completely reproducible and must compile. No late homeworks will be accepted.

Reading Read the paper Binette and Steorts (2022) to get an overview of entity resolution. You'll want to refer to this during the course of the semester as it's meant to be a quick reference regarding the concepts that we will be covering. For more details, refer to the book by Christen (2012).

Advice: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Total points on assignment: 5 (reproducibility) + 30 points for the assignment.

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. Perform an exploratory data analysis of the data set. Illustrate your findings with tables, visualizations and provide commentary of your findings.

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.

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

Hint 2: Post a question if you're getting stuck. This is a hard but very important problem. Do your best to work through this on your own.