CS 208 - Applied Privacy for Data Science Homework 1

Spring 2019
Harvard University

Problem 1

Studying Latanya Sweeney's record linkage reidentification attack, she used `zip code`, `date of birth`, and `gender` to uniquely identify a large portion of the population.

The dataset was loaded into `R`, where preliminary data exploration took place. Most notably, there are 1000 entries and 22 variables in the dataset, with the variables being:

`X.1, state, puma, X, jpumarow, serialno.household, sex, age, educ, income, latino, black, asian, married, divorced, uscitizen, children, disability, militaryservice, employed, englishability, fips`

The naive and most effective starting point is tallying the unique values for each of these variables. Two of the variables, `state` and `fips`, have the same value for all 1000 rows and therefore will not be considered at all. On the opposite side of the spectrum, `X.1` is the unique ID and therefore will be disregarded.

Examining location identifiers, the question implies the inclusion of `puma`. Examining the rest of the variables, note that `jpumarow` = `puma` + 1090, meaning that it provides absolutely no unique information beyond what is already known from `puma` and can thus be ignored. Furthermore, here is a summary of counts for each PUMA value.

PUMA	1101	1102	1103	1104	1105	1106	1107
Count	117	241	140	154	116	127	105

These are very roughly equal (i.e. on the same order of magnitude).

Problem 2

Appendix