Demography SpawnR

R package that is intended to take an input dataset and be able to simulate a new dataset based on the distributions of the variables in the input dataset.

Reads each variable 1 at a time and determines its type: categorical (how many levels?), continuous (rounded/decimal), addresses/names/sensitive information, dates that are in the “*yyyymmdd*,” or the user is able to specify the format of dates (also possible to allow for multiple date formats – still working on this. Proving to be a little trickier than before).

After reading the variable type, the idea is to be able to take the distributions of those variables and output a dataset that is virtually the same as the original dataset but with PII removed. For example, if a variable is a categorical variable, we compute the frequencies/percentages of each level (including missing and NA values) and use those as sampling weights in the output simulated dataset. If a variable is continuous, for simplicity, we assume a normal distribution (knowing fully well this is not always the case) with the mean equal to the mean of the variable in the input dataset, and respective standard deviation (excluding missing and NA values). If missing data are present, we calculate the proportion of missing/NA values and after populating the entire column with values from a normal distribution, we either keep the value or insert a missing/NA into each row with a probability of [the proportion of column that is missing]. Lastly, if a variable is recognized to have

Another functionality is the ability to list all possible variable combinations that may be correlated or associated – continuous with continuous and categorical with categorical. Moreover, the user is able to calculate the respective correlations and p-values for each pair of variables. In addition to getting all variable pairs and their correlations, a user is also able to input categorical variables that they already know to be correlated and sample into the output dataset given the bivariate distribution of those 2 variables.

In addition to different types of categorical and continuous variables, ideally Demography SpawnR is able to recognize any standard date format and sample random dates given the range of the input dates. However, that is proving extremely tricky because manipulating dates in *R* is hard.

The functions are scalable up-to virtually any size input and output dataset, but the runtime depends on how patient you are ☺