Successful Synthetic Dataset Generation from Python pgmpy Package

This summarizes the **successful synthetic dataset generation** of DEP State of the Community Survey data using python's pgmpy package. Recall that the objective is to create a model from from the original dataset for synthetic data generation, to protect privacy. This attempt was successful in meeting three output criteria:

- 1. No duplicates of synthetic vs. original dataset to protect privacy
- 2. Distribution of all columns of synthetic should match original distributions
- 3. Quick check: Salary split by Education status should match

OUTCOMES: All technical and output criteria met.

- Bayesian Network Model used here has 70 arcs and zero duplicates vs. Original dataset for all 5 runs of parameter tuning. (Attachment 1 – Figure 1). Other attempts also successful, but this model has the lowest arcs and lowest prior weights while meeting 'no isolated nodes' technical criteria.
- 2. Distribution of synthetic dataset matches original dataset. (See accompanying pdf: Distribution of Original vs. Synthetic Datasets.pdf)
- 3. Further split of salary ranges by latest education status is also comparable to original dataset. (Attachment 1 -Figure 2)

NEXT STEPS:

Document the results, the workflow process (Attachment 2) and technical details (e.g. parameter tuning) in github and prepare synthetic dataset for release.

Prepared by:

Sandy G. Cabanes

Data Engineering Pilipinas - Moderator

Attachment 1 Bayesian network model from python's pgmpy vs. R's bnlearn

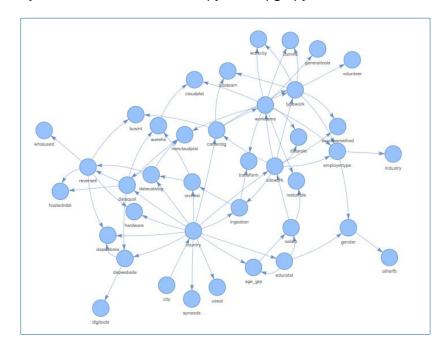


Figure 1 - Bayesian network from **PYTHON's** pgmpy ess = 1500

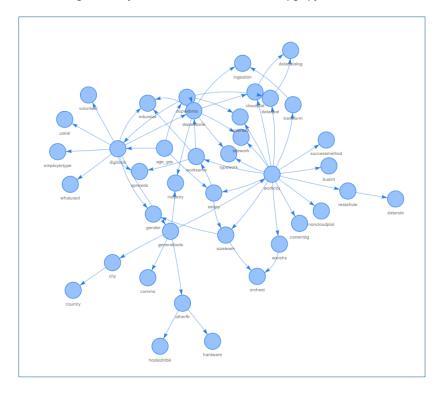


Figure 2 - Bayesian network from **R's bnlearn**

Attachment 2: Workflow flowchart

SYNTHETIC DATASET GENERATION USING PYTHON PGMPY PACKAGE FOR SURVEY DATA Uses : Bayesian Network, HillClimbSearch, BDeu (Network model) BayesianEstimator, MaximumLikelihoodEstimator (CPD) Bayesian Model Sampling (Synth generation) -START-Import packages Pre-process data -PARAMETER TUNING-Global NB starter ESS range (priors) Global ek specify Loop ESS sweep With isolated nodes Isolated nodes? + *Loop to: ESS range No isolated n → Fit CPD Bayesian Estimator (OR Max Likelihood Est) -PRIVACY CHECK---Synthetic data With duplicates Duplicates? vs. Original *Loop to: Fit CPD No duplicates Original distributions Distributions match? not matched vs. Original *Loop to: Fit CPD SANITY CHECKS — Salary by Educstat Distributions match not match *Loop to: Fit CPD Salary by Educstat Salary by Educ ok Synth dataset final ~~~~~~ Synthetic dataset processing ---- END -

Originated by: Sandy G. Cabanes Beyond surveys. Data-Powered Decisions.

Draft Flowchart app: https://github.com/SandyGCabanes/Unicode-Flowchart-Builder-App