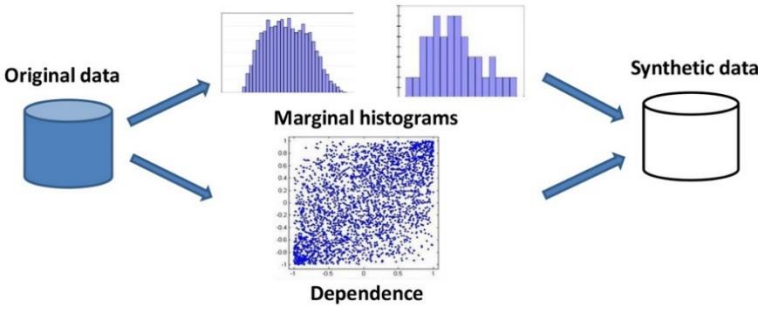
Synthetic Data Vault 0.13.0

SDV The Synthetic Data Vault	**Synthetic Data**	
	Systeemvereisten:	Python 3.6+
	Ontwikkeld door:	MIT Data to AI lab, Datacebo Inc., https://sdv.dev/
MIT License	Contactpersoon:	joachim.ganseman@smals.be

Functionaliteiten

Synthetic Data Vault is een open source Python library voor het aanmaken van synthetische data (zie ook https://www.smalsresearch.be/synthetic-data/). Vertrekkend van een gegeven dataset, in de vorm van een tabel, probeert ze de distributies en de correlaties in een dataset aan te leren door met statistische of deep learning – technieken een generatief model te trainen op die gegeven dataset. Dat model kan dan gebruikt worden om ongelimiteerd nieuwe gegevens aan te maken, die bij benadering hetzelfde formaat en dezelfde statistische eigenschappen aanhouden (op voorwaarde van succesvolle training).

De library voorziet basisfunctionaliteit om eenvoudige constraints op te leggen (<, >, =, etc.), om subsets te maken door bij het genereren enkele variabelen vast te zetten op een bepaalde waarde, en om type, distributie en bereik van variabelen desgewenst manueel te specifiëren. Aparte routines zijn voorzien voor time series en relationele data. De zusterlibraries *SDMetrics* en *SDGym* kunnen gebruikt worden om de gegenereerde dataset te evalueren volgens statistische of andere metrieken.



(afbeelding uit Haoran Li et al: *DPSynthesizer: Differentially Private Data Synthesizer for Privacy Preserving Data Sharing*)

Conclusies & Aanbevelingen

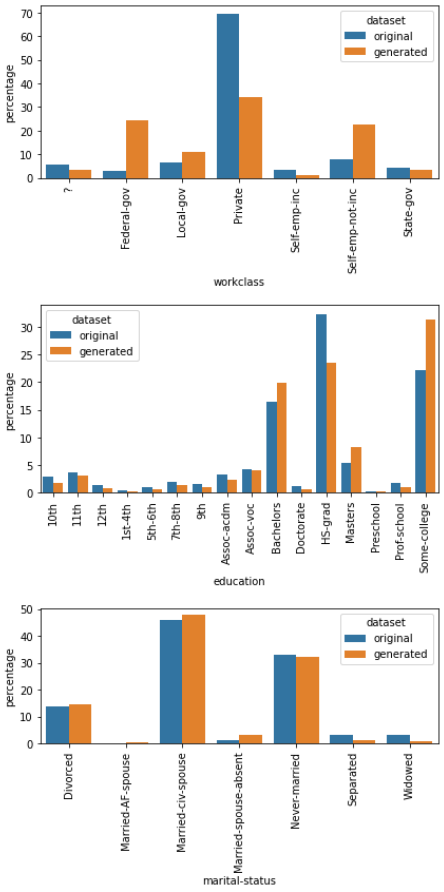
SDV is een populaire open source library voor het genereren van een nieuwe synthetische dataset naar het model van een zelf opgegeven dataset. Ze maakt daarbij gebruik van de recente technieken waaronder deep learning. Ze is gemakkelijk te integreren in een data processing workflow op basis van Python. De library is nog in alfa-status en actief in ontwikkeling, wat vandaag nog een zekere uitdaging vormt qua onderhoudbaarheid en stabiliteit. Bij sommige operaties kan men nog op bugs botsen.

Testen & Resultaten

We testten deze library uitgebreid bij een project voor een van onze klanten om gegevens te “scramblen” ter bescherming van de privacy. De basisfunctionaliteit van de library is gemakkelijk aan te wenden, en de voorbeelden uit de documentatie zijn snel aanpasbaar om ze te kunnen toepassen op eigen gegevens.

Van de vier beschikbare methodes om data te genereren is de statistische “Gaussian Copula” methode de eenvoudigste. De alternatieve methodes “CopulaGan”, “CTGan”, “TVAE”, gebruiken *PyTorch* in de achtergrond, zijn multi-threaded en kunnen eventueel grafische kaarten benutten, maar op tabulaire en tekstuele gegevens merken we daarmee in de praktijk nauwelijks versnelling. Bij onze tests verkregen we de meest kwalitatieve resultaten met de “CopulaGAN” methode.

De library werkt het best met tabellen met weinig kolommen (variabelen), en voor elke waarde van elke variabele veel rijen. Zo zijn er voldoende gegevens om correcte partiële distributies aan te leren en een goed model te bekomen. Als distributies erg extreem zijn (bepaalde waarden zijn erg zeldzaam), merken we een groot risico op overfitting. Dan kunnen bij het genereren van gegevens, erg zeldzame waarden plots helemaal verdwijnen. Dit kan deels verholpen worden door te controleren of de distributies en datatypes wel correct zijn gedetecteerd, en ze desgewenst te forceren naar andere opties – wat wel enige achtergrondkennis statistiek vergt.



SDV is een jong project en kampt nog met onvolledigheden en bugs. Zo merken we ongedocumenteerde interne formaatconversies, problemen bij ontbrekende data, gebrekkige integratie met andere libraries zoals Faker, en onvolledige documentatie. Door te vertrekken van op voorhand opgeschoonde data, en zelf de nodige verificaties toe te voegen, kunnen de meeste van deze beperkingen snel omzeild worden.

SDV is enkel een library. Visualisaties van het resultaat, zoals afgebeeld hiernaast, moeten desgewenst afzonderlijk bijgeprogrammeerd worden. Tot slot, de nuttigste truc om kwaliteit en snelheid te verhogen, blijkt in onze ervaring te zijn om het aantal kolommen waarvoor een generatief model wordt getraind, te beperken tot het minimum dat noodzakelijk is om het beoogde doel te bereiken.

Gebruiksvoorwaarden & Budget

Synthetic Data Vault is een open source library en is gratis en vrij beschikbaar. De code staat op Github onder de zeer permissieve MIT licensie. Rond het initiële ontwikkelteam van SDV is recent spinoff-bedrijf Datacebo gevormd, dat commerciële ondersteuning kan bieden.