Synthetic Data Generation

Review and Solutions

Agis Kounelis

kounelisagis@gmail.com

May 2021

Synthetic Data

- **Why?**
  Access to large enough datasets with real data (such as Citizen ID, Birth Certificate, Passport, Health Insurance, Address, Medical History, etc.) is not nearly as common as access to toy datasets on Kaggle, specifically designed for machine learning tasks. Thus, we need a few lines of code to generate large datasets with random meaningful entries.

Synthetic Data

- **What?**
  - It can be numerical, binary, or categorical
  - The number of features and length of the dataset should be arbitrary
  - It should preferably be random and the user should be able to choose a wide variety of statistical distributions to base this data upon
  - Random noise can be interjected in a controllable manner

Solutions

Online:

- Mockaroo (Paid): mockaroo.com

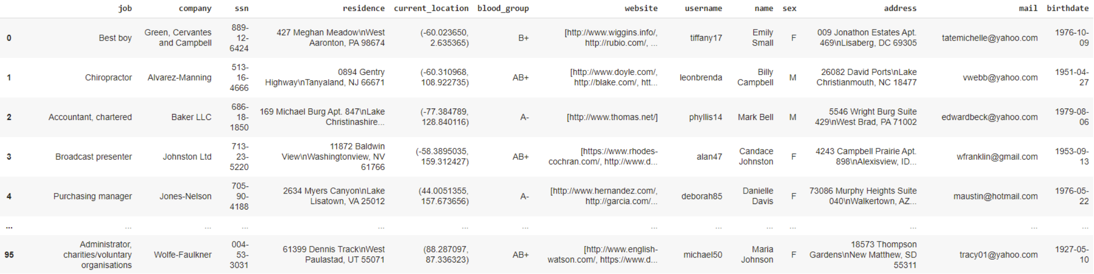
Nodejs:

1. Randomuser.me: github.com/RandomAPI/Randomuser.me-Node
2. json-schema-faker: github.com/json-schema-faker/json-schema-faker

Python:

1. Faker: github.com/joke2k/faker
2. Mimesis: github.com/lk-geimfari/mimesis

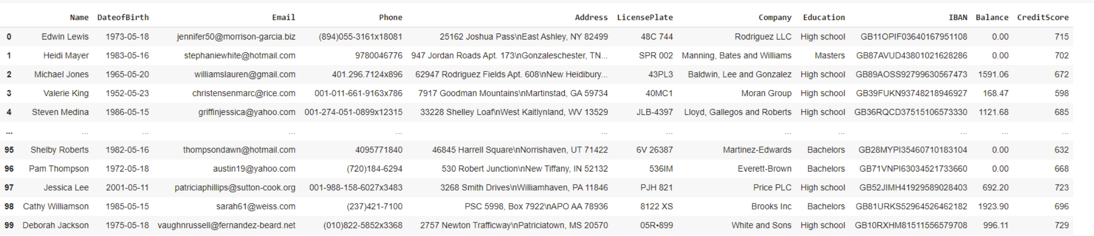
Faker Example (1)



	job	company	ssn	residence	current_location	blood_group	website	username	name	sex	address	mail	birthdate
0	Best boy	Green, Cervantes and Campbell	889-12-6424	427 Meghan Meadow\nWest Aaronton, PA 98674	(-60.023650, 2.635365)	B+	[http://www.wiggins.info/, http://rubio.com/, ...	tiffany17	Emily Small	F	009 Jonathon Estates Apt. 469\nLisaberg, DC 69305	tatemichelle@yahoo.com	1976-10-09
1	Chiropractor	Alvarez-Manning	513-16-4666	0894 Gentry Highway\nTanyaland, NJ 66671	(-60.310968, 108.922735)	AB+	[http://www.doyle.com/, http://blake.com/, htt...	leonbrenda	Billy Campbell	M	26082 David Ports\nLake Christianmouth, NC 18477	vwebb@yahoo.com	1951-04-27
2	Accountant, chartered	Baker LLC	686-18-1850	169 Michael Burg Apt. 847\nLake Christinashire...	(-77.384789, 128.840116)	A-	[http://www.thomas.net/]	phyllis14	Mark Bell	M	5546 Wright Burg Suite 429\nWest Brad, PA 71002	edwardbeck@yahoo.com	1979-08-06
3	Broadcast presenter	Johnston Ltd	713-23-5220	11872 Baldwin View\nWashingtonview, NV 61766	(-58.3895035, 159.312427)	AB+	[https://www.rhodes-cochran.com/, http://www.d...	alan47	Candace Johnston	F	4243 Campbell Prairie Apt. 898\nAlexisview, ID...	wfranklin@gmail.com	1953-09-13
4	Purchasing manager	Jones-Nelson	705-90-4188	2634 Myers Canyon\nLake Lisatown, VA 25012	(44.0051355, 157.673656)	A-	[http://www.hernandez.com/, http://garcia.com/...	deborah85	Danielle Davis	F	73086 Murphy Heights Suite 040\nWalkertown, AZ...	maustin@hotmail.com	1976-05-22
...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	Administrator, charities/voluntary organisations	Wolfe-Faulkner	004-53-3031	61399 Dennis Track\nWest Paulastad, UT 55071	(88.287097, 87.336323)	AB+	[http://www.english-watson.com/, https://www.d...	michael50	Maria Johnson	F	18573 Thompson Gardens\nNew Matthew, SD 55311	tracy01@yahoo.com	1927-05-10

Figure: Profiles using the built-in profile(.) function

Faker Example (2)



	Name	DateofBirth	Email	Phone	Address	LicensePlate	Company	Education	IBAN	Balance	CreditScore
0	Edwin Lewis	1973-05-18	jennifer50@morrison-garcia.biz	(894)055-3161x18081	25162 Joshua Pass\nEast Ashley, NY 82499	48C 744	Rodriguez LLC	High school	GB11OPIF03640167951108	0.00	715
1	Heidi Mayer	1983-05-16	stephaniewhite@hotmail.com	9780046776	947 Jordan Roads Apt. 173\nGonzaleschester, TN...	SPR 002	Manning, Bates and Williams	Masters	GB87AVUD43801021628286	0.00	702
2	Michael Jones	1965-05-20	williamslauren@gmail.com	401.296.7124x896	62947 Rodriguez Fields Apt. 608\nNew Heidibury...	43PL3	Baldwin, Lee and Gonzalez	High school	GB89AOSS92799630567473	1591.06	672
3	Valerie King	1952-05-23	christensenmarc@rice.com	001-011-661-9163x786	7917 Goodman Mountains\nMartinstad, GA 59734	40MC1	Moran Group	High school	GB39FUKN93748218946927	168.47	598
4	Steven Medina	1986-05-15	griffinjessica@yahoo.com	001-274-051-0899x12315	33228 Shelley Loaf\nWest Kaitlynland, WV 13529	JLB-4397	Lloyd, Gallegos and Roberts	High school	GB36RQCD37515106573330	1121.68	685
...	...	...	...	...	...	...	...	...	...	...	...
95	Shelby Roberts	1982-05-16	thompsondawn@hotmail.com	4095771840	46845 Harrell Square\nNorrishaven, UT 71422	6V 26387	Martinez-Edwards	Bachelors	GB28MYPI35460710183104	0.00	632
96	Pam Thompson	1972-05-18	austin19@yahoo.com	(720)184-6294	530 Robert Junction\nNew Tiffany, IN 52132	536IM	Everett-Brown	Bachelors	GB71VNPI63034521733660	0.00	668
97	Jessica Lee	2001-05-11	patriciaphillips@sutton-cook.org	001-988-158-6027x3483	3268 Smith Drives\nWilliamhaven, PA 11846	PJH 821	Price PLC	High school	GB52JIMH41929589028403	692.20	723
98	Cathy Williamson	1985-05-15	sarah61@weiss.com	(237)421-7100	PSC 5998, Box 7922\nAPO AA 78936	8122 XS	Brooks Inc	Bachelors	GB81URKS52964526462182	1923.90	696
99	Deborah Jackson	1975-05-18	vaughnrussell@fernandez-beard.net	(010)822-5852x3368	2757 Newton Trafficway\nPatriciatown, MS 20570	05R•899	White and Sons	High school	GB10RXHM81511556579708	996.11	729

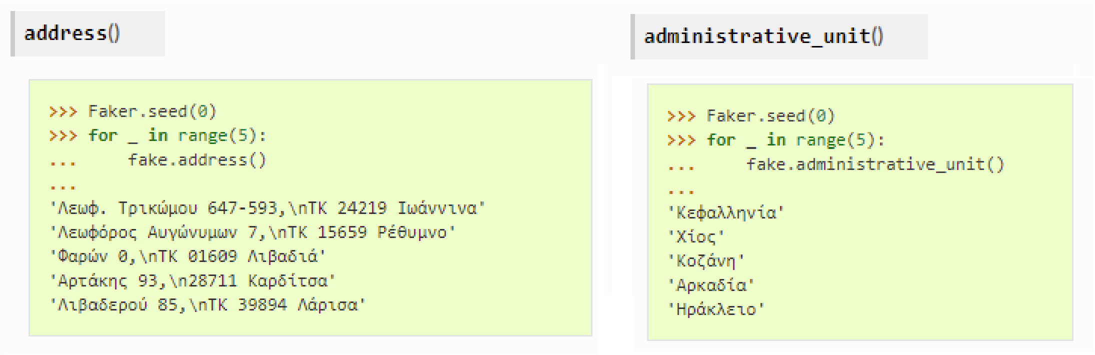
Figure: Profiles using the Standard Providers of the library and numpy functions

Why choosing a more complicated solution?

More control over the distribution of some fields. For example a numeric field (Balance) can follow a Poisson distribution.

Faker

There are options for Localized Providers



address()

```
>>> Faker.seed(0)
>>> for _ in range(5):
...     fake.address()
...
'Λεωφ. Τρικώμου 647-593,\nΤΚ 24219 Ιωάννινα'
'Λεωφόρος Αυγώνυμων 7,\nΤΚ 15659 Ρέθυμνο'
'Φαρών 0,\nΤΚ 01609 Λιβαδιά'
'Αρτάκης 93,\n28711 Καρδίτσα'
'Λιβαδερού 85,\nΤΚ 39894 Λάρισα'
```

administrative_unit()

```
>>> Faker.seed(0)
>>> for _ in range(5):
...     fake.administrative_unit()
...
'Κεφαλληνία'
'Χίος'
'Κοζάνη'
'Αρκαδία'
'Ηράκλειο'
```

Figure: Profiles using the Locale el_GR option

Next Steps

- Specification of data fields
- Creation of datasets specifically for countries of interest