DataEng: Data Ethics In-class Assignment

Chase Verbout

This week you will use various techniques to construct synthetic data.

Submit: Make a copy of this document and use it to record your responses and results (use colored highlighting when recording your responses/results). Store a PDF copy of the document in your git repository along with your code before submitting for this week.

A. [MUST] Discussion Questions

A ride-share company (similar to Lyft or Uber) decides to publish detailed ride data to encourage researchers to develop ideas and open source software that might someday enhance the company's products. The company's data engineer publishes the complete set of ride trips for a single year. Data for each trip includes start location, end location, GPS breadcrumb data during trip, price charged, mileage, number of riders served, and information about make, model and year of the vehicle that serviced the trip. All personal information (names, ages, addresses, birthdates, account information, payment information, credit card numbers, etc.) is stripped from the data before sharing.

Can you see a problem with this approach? How might an attacker re-identify some of the real passengers? Insert your responses here and discuss with your group members.

I definitely see the potential for problems in this approach. For instance, geographic information can be fairly revealing especially when combined with timestamps. If we know where someone was and when they were there, we've narrowed down the data quite quickly. Sometimes people repeatedly make the same trip- this would be identifiable here. And lastly, trips to sensitive locations are now available which can be identifiable because of publicly available information.

Search the internet and provide a URL of one article that describes one data breach that occurred during the previous 5 years. The breach must be one in which the attacker obtained personal, private information about customers or employees of the attacked enterprise.

Biggest Data Breaches in US History (Updated 2024) | UpGuard

Briefly summarize the breach here, Which of the techniques discussed in the lecture might help to prevent this sort of problem in the future? Describe your chosen breach and your recommendations with your group members.

In January of 2021, Microsoft Exchange was breached impacting 30,000 US companies and 60,000 companies internationally. Gaining access to emails from small businesses and local governments, the hackers were able to take control of vulnerable systems deploying malware and accessing sensitive data.

The hackers did so through 4 zero-day vulnerabilities wherein they only needed connection to the internet and on premise, locally managed systems.

Of the cybersecurity measures we discussed I think that the most beneficial aspects would have been encryption on the data such that it may have been useless without the key. Employee training and awareness to better understand the importance of actions taken and what is suspicious.

B. [SHOULD] Sampling

Use the DataFrame sample() method to produce a 20 element sample of the data. Use the “weights” parameter of the sample() method to synthetically bias the sample such that employees with ages 40-49 are three times as likely to be sampled as employees in other age ranges.

C. [SHOULD] Anonymization

Anonymize the name (both first and last names), email, and phone number information in the employee data.

D. [SHOULD] Perturbation

Perturb the age, salary and years of experience attributes of the employees data using Gaussian noise. How should we choose the standard deviation parameter for the noise? Should we choose the same standard deviation for all three of the perturbed attributes? If not, then how should we choose?

E. [MUST] Model Based Synthesis

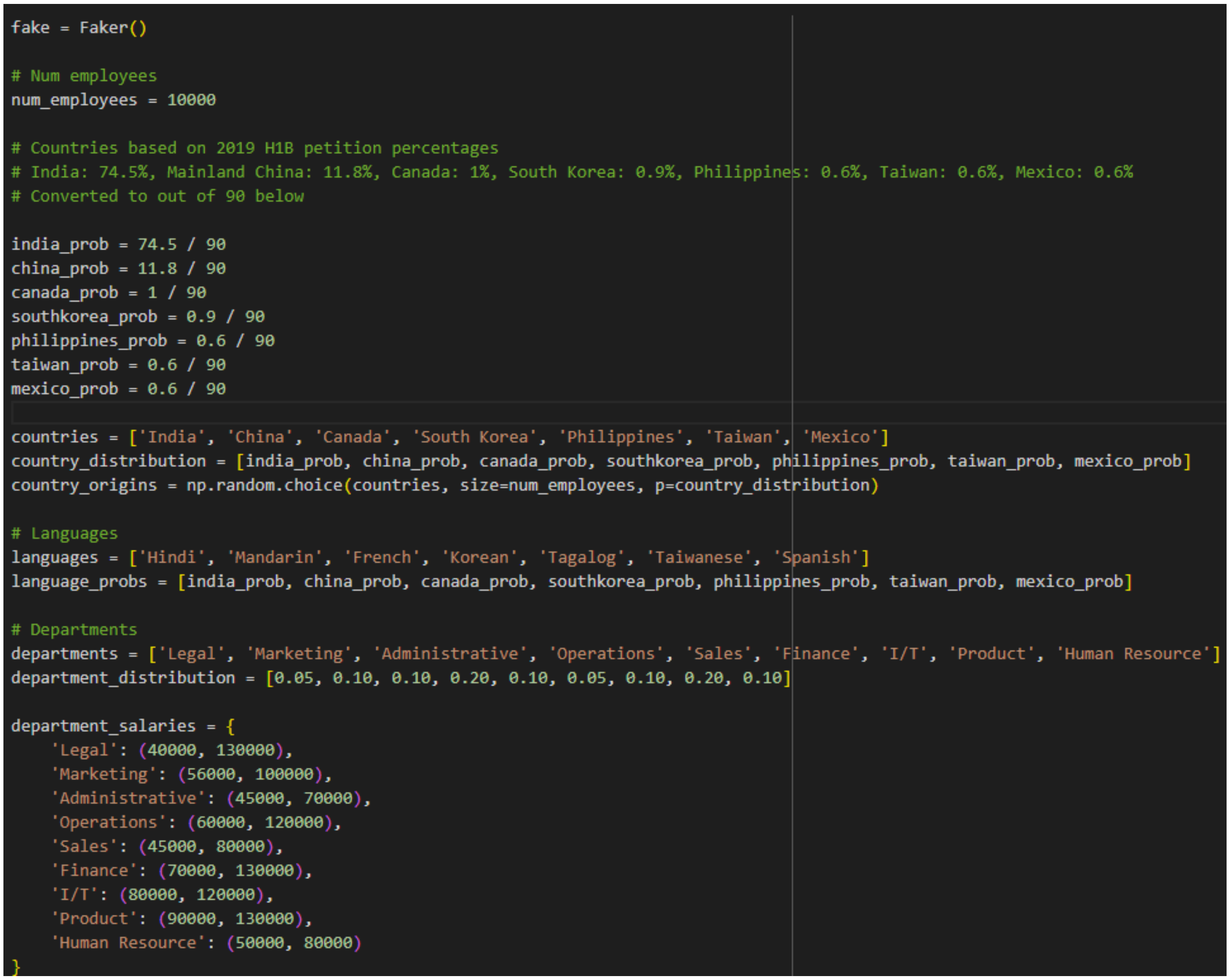
Your job is to synthesize a data set based on the employees.csv data set

This startup company of 320 employees intends to go public and become a 10,000 employee company. Your job is to produce an expanded 10K record synthetic database to help the founders understand personnel-related issues that might occur with the expanded company.

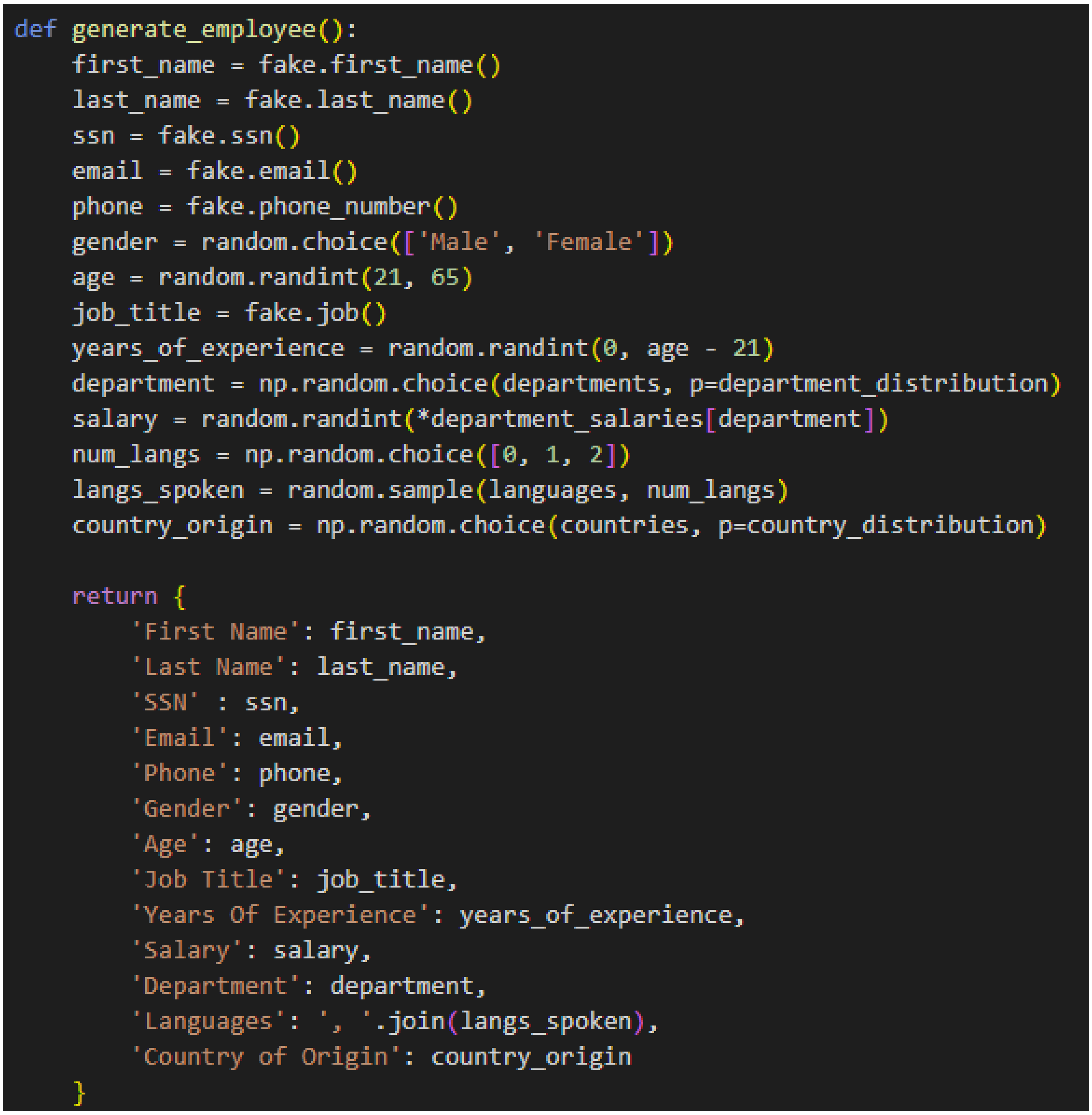
Use the Faker python module to produce a 10K employee dataset. Follow these constraints:

- All columns in the current data set must be preserved. It is not necessary to preserve any of the actual data from the current database
- Need to keep track of social security numbers
- The database should keep track of the languages (other than English) spoken by each employee. Each employee speaks 0, 1 or 2 languages in addition to English.
- To grow, the company plans to sponsor visas and hire non-USA citizens. So your synthetic database should include names of employees from India, Mainland China, Canada, South Korea, Philippines, Taiwan and Mexico. These names should be in proportion to the 2019 percentages of H1B petitions from each country.
- The expanded company will have additional departments include “Legal” (approximately 5% of employees), “Marketing” (10%), “Administrative” (10%), “Operations” (20%), “Sales” (10%), “Finance” (5%) and “I/T” (10%) to go along with the current “Product” (20%) and “Human Resource” (10%) departments.
- Salaries in each department must mimic the typical salaries for professionals in each field. You can find appropriate data for each type of profession at salary.com For example, see this page to find a model estimate for your synthetic marketing department: https://www.salary.com/research/salary/benchmark/marketing-specialist-salary
- The current startup company (as represented by the employees.csv data) is skewed toward male employees. Our goal for the new company is to make the numbers of men and women approximately equal.

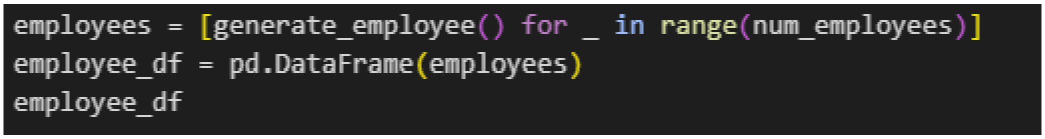
Save your new database to your repository alongside your code that synthesized the data.



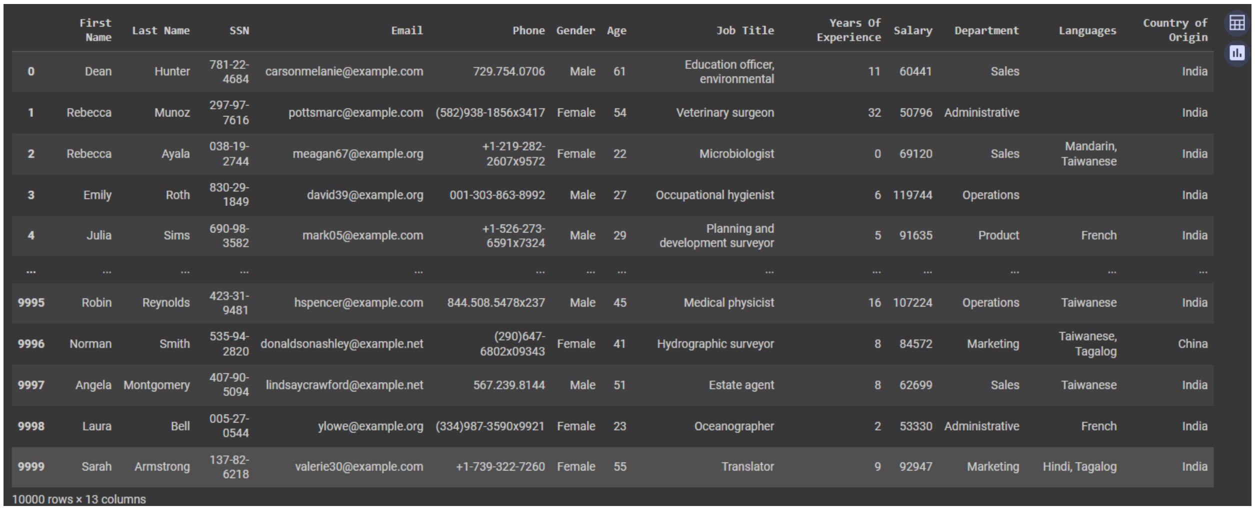
```python
fake = Faker()

# Num employees
num_employees = 10000

# Countries based on 2019 H1B petition percentages
# India: 74.5%, Mainland China: 11.8%, Canada: 1%, South Korea: 0.9%, Philippines: 0.6%, Taiwan: 0.6%, Mexico: 0.6%
# Converted to out of 90 below

india_prob = 74.5 / 90
china_prob = 11.8 / 90
canada_prob = 1 / 90
southkorea_prob = 0.9 / 90
philippines_prob = 0.6 / 90
taiwan_prob = 0.6 / 90
mexico_prob = 0.6 / 90

countries = ['India', 'China', 'Canada', 'South Korea', 'Philippines', 'Taiwan', 'Mexico']
country_distribution = [india_prob, china_prob, canada_prob, southkorea_prob, philippines_prob, taiwan_prob, mexico_prob]
country_origins = np.random.choice(countries, size=num_employees, p=country_distribution)

# Languages
languages = ['Hindi', 'Mandarin', 'French', 'Korean', 'Tagalog', 'Taiwanese', 'Spanish']
language_probs = [india_prob, china_prob, canada_prob, southkorea_prob, philippines_prob, taiwan_prob, mexico_prob]

# Departments
departments = ['Legal', 'Marketing', 'Administrative', 'Operations', 'Sales', 'Finance', 'I/T', 'Product', 'Human Resource']
department_distribution = [0.05, 0.10, 0.10, 0.20, 0.10, 0.05, 0.10, 0.20, 0.10]

department_salaries = {
    'Legal': (40000, 130000),
    'Marketing': (56000, 100000),
    'Administrative': (45000, 70000),
    'Operations': (60000, 120000),
    'Sales': (45000, 80000),
    'Finance': (70000, 130000),
    'I/T': (80000, 120000),
    'Product': (90000, 130000),
    'Human Resource': (50000, 80000)
}
```



```
def generate_employee():
    first_name = fake.first_name()
    last_name = fake.last_name()
    ssn = fake.ssn()
    email = fake.email()
    phone = fake.phone_number()
    gender = random.choice(['Male', 'Female'])
    age = random.randint(21, 65)
    job_title = fake.job()
    years_of_experience = random.randint(0, age - 21)
    department = np.random.choice(departments, p=department_distribution)
    salary = random.randint(*department_salaries[department])
    num_langs = np.random.choice([0, 1, 2])
    langs_spoken = random.sample(languages, num_langs)
    country_origin = np.random.choice(countries, p=country_distribution)

    return {
        'First Name': first_name,
        'Last Name': last_name,
        'SSN' : ssn,
        'Email': email,
        'Phone': phone,
        'Gender': gender,
        'Age': age,
        'Job Title': job_title,
        'Years Of Experience': years_of_experience,
        'Salary': salary,
        'Department': department,
        'Languages': ', '.join(langs_spoken),
        'Country of Origin': country_origin
    }
```



```
employees = [generate_employee() for _ in range(num_employees)]
employee_df = pd.DataFrame(employees)
employee_df
```



	First Name	Last Name	SSN	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department	Languages	Country of Origin
0	Dean	Hunter	781-22-4684	carsonmelanie@example.com	729.754.0706	Male	61	Education officer, environmental	11	60441	Sales		India
1	Rebecca	Munoz	297-97-7616	pottsmarc@example.com	(582)938-1856x3417	Female	54	Veterinary surgeon	32	50796	Administrative		India
2	Rebecca	Ayala	038-19-2744	meagan67@example.org	+1-219-282-2607x9572	Female	22	Microbiologist	0	69120	Sales	Mandarin, Taiwanese	India
3	Emily	Roth	830-29-1849	david39@example.org	001-303-863-8992	Male	27	Occupational hygienist	6	119744	Operations		India
4	Julia	Sims	690-98-3582	mark05@example.com	+1-526-273-6591x7324	Male	29	Planning and development surveyor	5	91635	Product	French	India
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	Robin	Reynolds	423-31-9481	hspencer@example.com	844.508.5478x237	Male	45	Medical physicist	16	107224	Operations	Taiwanese	India
9996	Norman	Smith	535-94-2820	donaldsonashley@example.net	(290)647-6802x09343	Female	41	Hydrographic surveyor	8	84572	Marketing	Taiwanese, Tagalog	China
9997	Angela	Montgomery	407-90-5094	lindsaycrawford@example.net	567.239.8144	Male	51	Estate agent	8	62699	Sales	Taiwanese	India
9998	Laura	Bell	005-27-0544	ylowe@example.org	(334)987-3590x9921	Female	23	Oceanographer	2	53330	Administrative	French	India
9999	Sarah	Armstrong	137-82-6218	valerie30@example.com	+1-739-322-7260	Female	55	Translator	9	92947	Marketing	Hindi, Tagalog	India

10000 rows × 13 columns

G. [SHOULD] Analyze the Synthetic Company

- How many men vs. women will we need to hire in each department?
- How much will this new company pay in yearly payroll?
- Other than hiring from non-USA countries, how else might the company grow quickly from size=320 to size=10000?
- How much office space will this company require?
- Does this new dataset preserve the privacy of the original employees listed in employees.csv?

H. [ASPIRE] Quality of the Synthetic Dataset

Use ydata-profiling to explore your synthetic data set: https://pypi.org/project/ydata-profiling/
Use ydata-profiling with the original employees.csv as well to compare.

In what ways does the synthetic data set appear to be obviously synthetic and/or not representative of the current company?

How might you improve the synthetic data to make it more realistic?