

DataEng: Data Ethics In-class Assignment

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.

There are following problems with this model:

Geographical data can be used to identify users' home and office locations.

Unique locations can be exposed if a user visits a distinctive place.

Hackers can combine various data sets to impersonate someone else.

Attackers can use external data sources to correlate with published ride data.

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.

<https://www.itgovernance.co.uk/blog/list-of-data-breaches-and-cyber-attacks-in-2023>

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 October 2019, LifeLabs, a Canadian lab testing company, experienced a significant data breach. This incident exposed the personal health information of 15 million individuals, including 85,000 test results from Ontario. The breach occurred due to inadequate security measures and excessive data collection, which heightened the risk of identity theft and financial issues for customers (Global News).

To address such issues, the following techniques can be implemented:

Regular Security Audits: Conduct frequent and comprehensive security assessments to identify and rectify vulnerabilities.

Data Encryption: Implement encryption for sensitive data during transmission and storage to protect it from unauthorized access.

Minimize Data Collection: Collect only the essential data needed to reduce the risk associated with storing excessive information.

Routine Data Deletion: Establish protocols for regularly deleting data that is no longer necessary.

Employee Cybersecurity Training: Provide ongoing training to employees to help them recognize and respond effectively to security threats.

Phishing Awareness Simulations: Regularly conduct phishing simulations to train employees on identifying and avoiding phishing attacks.

Incident Response Planning and Drills: Develop a comprehensive incident response plan and conduct regular drills to ensure preparedness in case of a data breach.

Multi-Factor Authentication (MFA): Use MFA for accessing sensitive systems to add an additional layer of security.

Implementing these strategies can help organizations significantly reduce the risk of data breaches and safeguard sensitive customer information.

B. [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 40% employees who are non-USA citizens and should include names of employees from India, Mainland China, Canada, South Korea,

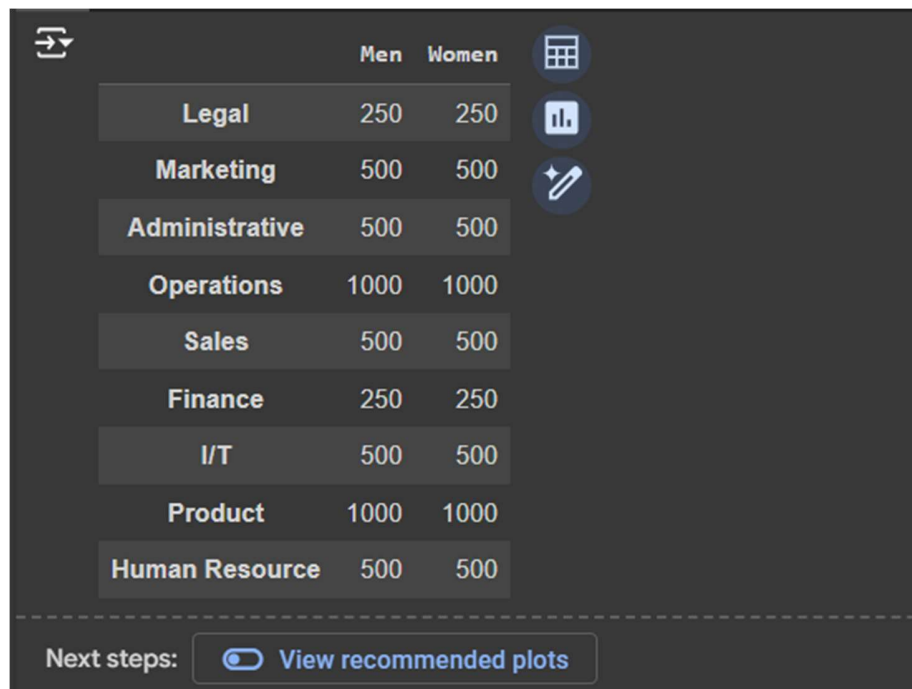
Philippines, Taiwan and Mexico. These names should be in proportion to [the 2019 percentages of H1B petitions from each country](#).

- 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.

C. [SHOULD] Analyze the Synthetic Company

- How many men vs. women will we need to hire in each department?



The screenshot shows a dashboard with a table of employee counts by department and gender. The table has columns for Department, Men, and Women. The departments listed are Legal, Marketing, Administrative, Operations, Sales, Finance, I/T, Product, and Human Resource. The counts for Men and Women are equal for each department. To the right of the table are three icons: a calendar, a bar chart, and a pencil. At the bottom, there is a 'Next steps:' section with a toggle switch and a button labeled 'View recommended plots'.

	Men	Women
Legal	250	250
Marketing	500	500
Administrative	500	500
Operations	1000	1000
Sales	500	500
Finance	250	250
I/T	500	500
Product	1000	1000
Human Resource	500	500

Next steps: ☐ View recommended plots

- How much will this new company pay in yearly payroll?

```
# Salary ranges for each department
SALARY_RANGES = {
    "Legal": (60000, 200000),
    "Marketing": (50000, 150000),
    "Administrative": (40000, 100000),
    "Operations": (50000, 150000),
    "Sales": (40000, 120000),
    "Finance": (50000, 150000),
    "I/T": (60000, 180000),
    "Product": (70000, 200000),
    "Human Resource": (50000, 120000)
}

# Calculate average salary for each department
average_salaries = {dept: (range_[0] + range_[1]) / 2 for dept, range_ in SALARY_RANGES.items()}

# Calculate total payroll for each department
total_payroll = {dept: count * average_salaries[dept] for dept, count in department_counts.items()}

# Calculate overall yearly payroll
yearly_payroll = sum(total_payroll.values())

yearly_payroll
```

104000000.0

- Other than hiring from non-US countries, how else might the company grow quickly from size=320 to size=10000?
Take smaller companies and merge them into one company, they can employ contract workers or freelancers which can help to scale up the workforce, invest in automation and technology, and remote work can help to expand it.
- How much office space will this company require? 1500000
- Does this new dataset preserve the privacy of the original employees listed in employees.csv?
The new dataset we have generated does not preserve the privacy of the original employees which are listed in employees.csv in case it has real employee data. We have personal data like SSNs, names, and contact data which is a problem to data privacy.
In order to ensure privacy:
 1. Anonymizing the data is required
 2. Need to mask the sensitive data called personally identifiable information PII
 3. Providing aggregated data instead of individual data

D. [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?

1. The synthesised data may have more uniform distribution of games, salaries, and experience levels in comparison to the original dataset.
2. We have fake email, and phone number pattern which is shows lack of diversity.
3. The distribution of language spoken may not match with the version real data.

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

1. Match the distribution of the age, experience and salary to that of original data
2. We can reflect actual jon titles from original data rather thank generating it from fake.
3. Ensure that the gender, age and department distribution matches the distribution of the original data
4. Using more realistic phone number and email of users
5. Using more accurate dara for language proficiency based on company's employees demographic

E. [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.

	First Name	Last Name	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department	Languages
9131	Amber	Yates	kim50@example.net	(967)750-5864x34039	Female	56.639526	Exercise physiologist	31.112685	48736.973173	Marketing	Korean
9545	Devin	Odom	tanya55@example.org	590.534.9410	Female	42.650362	Engineer, electrical	0.000000	74951.473785	I/T	
7390	Monica	Johnson	crobertson@example.org	5923168767	Male	41.403842	IT consultant	7.818812	117444.321594	Sales	
8237	Steven	Davis	angel69@example.org	648.413.9090	Female	24.360544	Consulting civil engineer	0.000000	151735.911964	Product	German
8468	David	Moore	qlawson@example.net	(413)479-4966x0928	Female	51.924440	Optician, dispensing	23.592516	108149.713953	Operations	
3372	Cheryl	Richards	alexis17@example.com	539-967-4357	Female	42.019027	Chief Financial Officer	2.219593	71588.774178	Human Resource	Mandarin
4024	Daniel	Lynch	victoriarmcdowell@example.org	312-771-6477	Female	33.422585	Youth worker	8.251156	141181.230556	Operations	
4778	Pamela	Garcia	elizabethltaylor@example.net	600-906-5047x60186	Female	36.825394	Gaffer	17.711752	103864.908121	Product	
212	Diane	Hall	uperez@example.net	(464)431-3185	Male	29.836534	Tax adviser	6.608164	51269.743801	Marketing	
7902	Sheri	Wright	tmiller@example.org	240-369-7739x72192	Female	58.894774	Local government officer	29.066707	109038.679493	Sales	
2242	Russell	Reynolds	jeremy56@example.org	+1-502-825-5000x314	Female	31.001482	Oceanographer	3.006572	84838.065236	Human Resource	Hindi, Korean
6626	Nathan	Gardner	sandersamanda@example.com	(516)764-5719	Male	35.833010	Commercial art gallery manager	7.944901	95799.299827	Operations	French
5696	Erik	Brooks	christopherbryant@example.net	372-561-9848x60041	Male	46.742047	Sports administrator	12.648760	67401.247327	Operations	Tagalog
6838	William	Gonzalez	bruce07@example.net	001-931-903-8505x654	Male	62.520162	Financial adviser	20.209723	143555.364987	Marketing	
5570	Evan	Bowen	vsmith@example.net	+1-644-856-9071	Female	33.517515	Volunteer coordinator	10.667596	122093.753146	I/T	
9086	Ashley	Rich	ejackson@example.net	428-584-5203x865	Female	23.502559	Occupational psychologist	0.000000	187002.453696	Product	
8577	Debra	Gutierrez	loriwilson@example.com	550-324-2810	Female	27.970684	Electrical engineer	11.148435	130563.356758	Operations	
249	Brandy	Walker	clinegary@example.org	9162054890	Male	49.120772	Engineer, building services	10.383677	70565.644157	Marketing	Tagalog
8359	Jesse	Love	pyoung@example.com	345.506.9607	Female	44.114736	Engineer, maintenance (IT)	0.000000	88752.920290	I/T	Korean
7308	Sara	Koch	travis34@example.org	001-953-577-7256	Female	39.844093	Environmental consultant	16.695490	122105.013201	Operations	

F. [SHOULD] Anonymization

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

	First Name	Last Name	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department	Languages	weight
4161	First Name _ 4161	Last Name_4161	Email_4161	Phone_4161	Male	48	Teacher, primary school	6	125702.68	Legal		3
7210	First Name _ 7210	Last Name_7210	Email_7210	Phone_7210	Female	53	Exercise physiologist	9	118460.86	Sales		1
0	First Name _ 0	Last Name_0	Email_0	Phone_0	Female	42	Mechanical engineer	15	147953.26	Operations	French	3
3007	First Name _ 3007	Last Name_3007	Email_3007	Phone_3007	Male	48	Colour technologist	14	127656.91	Marketing	Spanish	3
1432	First Name _ 1432	Last Name_1432	Email_1432	Phone_1432	Male	38	Radio broadcast assistant	5	125675.03	I/T		1
896	First Name _ 896	Last Name_896	Email_896	Phone_896	Male	63	Designer, graphic	20	146005.25	Operations		1
1836	First Name _ 1836	Last Name_1836	Email_1836	Phone_1836	Female	43	Engineer, manufacturing	17	114072.68	Human Resource	Hindi	3
3435	First Name _ 3435	Last Name_3435	Email_3435	Phone_3435	Female	39	Musician	11	81145.69	Marketing		1
3956	First Name _ 3956	Last Name_3956	Email_3956	Phone_3956	Male	49	Research officer, government	2	171602.97	Product		3

G. [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?

	First Name	Last Name	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department	Languages	weight
4161	First Name _ 4161	Last Name_4161	Email_4161	Phone_4161	Male	44	Teacher, primary school	1	112742.112503	Legal		3
7210	First Name _ 7210	Last Name_7210	Email_7210	Phone_7210	Female	53	Exercise physiologist	1	123381.805596	Sales		1
0	First Name _ 0	Last Name_0	Email_0	Phone_0	Female	44	Mechanical engineer	16	83507.217937	Operations	French	3
3007	First Name _ 3007	Last Name_3007	Email_3007	Phone_3007	Male	52	Colour technologist	15	144798.815585	Marketing	Spanish	3
1432	First Name _ 1432	Last Name_1432	Email_1432	Phone_1432	Male	40	Radio broadcast assistant	11	114393.589875	I/T		1
896	First Name _ 896	Last Name_896	Email_896	Phone_896	Male	65	Designer, graphic	17	146596.796995	Operations		1
1836	First Name _ 1836	Last Name_1836	Email_1836	Phone_1836	Female	41	Engineer, manufacturing	21	122989.249932	Human Resource	Hindi	3
3435	First Name _ 3435	Last Name_3435	Email_3435	Phone_3435	Female	40	Musician	25	54387.442539	Marketing		1

1. Age: When choosing the standard deviation for adjusting ages, consider the typical age range and how much variation exists in the original dataset. If the dataset includes a broad spectrum of ages, a larger standard deviation might be appropriate to introduce more variation. Conversely, if the age range is narrow, a smaller standard deviation should be enough. Generally, a standard deviation between 2 and 5 years is a sensible choice.

2. Salary: When determining the standard deviation for adjusting salaries, consider the distribution and variability of salaries in your dataset. If salaries span a wide range and there's a significant difference among employees' pay, a larger standard deviation is suitable. However, if salaries are fairly uniform, a smaller standard deviation will suffice. Typically, choosing a standard deviation between 5% and 20% of the average salary is reasonable.
3. Years of Experience: The standard deviation for adjusting years of experience should reflect the variability in the dataset. If there is a wide range of experience levels among employees, a larger standard deviation is appropriate. However, if most employees have similar years of experience, a smaller standard deviation will be sufficient. Generally, a standard deviation between 1 and 3 years is suitable for this purpose.

Choice of Standard Deviation:

You don't have to use the same standard deviation for every attribute when perturbing data. Instead, you can adjust the standard deviation to fit the unique characteristics and variability of each attribute in your dataset.

For example, if the age range in your data is broader and shows more variation compared to the salary range, you might select a larger standard deviation for modifying age. On the other hand, if years of experience show less variation than age and salary, you would use a smaller standard deviation for perturbing years of experience. This approach allows for more precise adjustments tailored to each attribute's specific variability.

In summary, the choice of standard deviation should be based on the characteristics of each attribute in the dataset, aiming to introduce realistic variation while preserving the overall distribution and characteristics of the original data.