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:

1. Geographical data can be taken to identify home and office locations of the users
2. Unique location can be reveled if the user visits unique place
3. An hacker can combine various data and fake it as other person
4. Attackers can take use of external data sources in order 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 company that does lab tests, had a big data breach. This incident revealed personal health information for 15 million people, including 85,000 test results from Ontario. The breach happened because their security was not strong enough and they collected too much data, putting customers at risk of identity theft and financial trouble (Global News).

Techniques to overcome the issue:

- **Regular Security Checks:** Perform frequent and thorough security checks to find and fix weaknesses.
- **Protect Data with Encryption:** Use encryption to secure sensitive data while it's being sent and when it's stored, keeping it safe from unauthorized access.
- **Limit Data Collection:** Gather just the essential data to minimize the risk of storing too much information.
- **Regular Data Purging:** Set up rules to regularly remove data that’s no longer necessary.
- **Cybersecurity Training:** Keep training employees regularly so they can identify and react to security threats effectively.
- **Phishing Simulations:** Run frequent phishing simulations to train employees on how to spot and avoid phishing attacks.
- **Develop and Test IR Plans:** Develop a thorough incident response plan and regularly practice drills to be ready in case of a data breach
- **MFA for Access:** Implement Multi-Factor Authentication (MFA) for accessing sensitive systems to provide an extra layer of security.

By adopting these techniques, organizations can significantly reduce the likelihood of data breaches and protect 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:

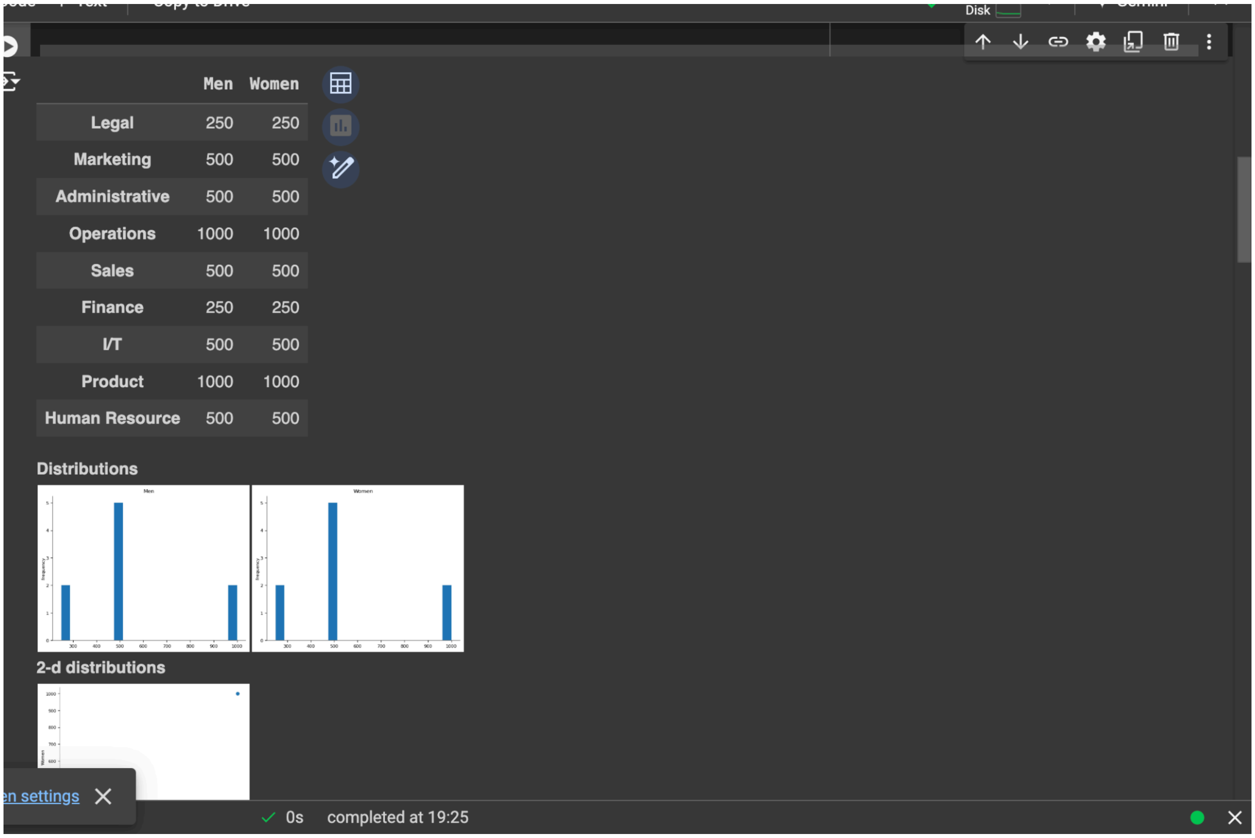
- All columns in the current data set must be preserved. It is not necessary to preserve any of the actual data from the current database
- Need to keep track of social security numbers

- 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?



	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

- How much will this new company pay in yearly payroll?1040000000.0
- Other than hiring from non-US countries, how else might the company grow quickly from size=320 to size=1000
  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?

  and) 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 liek SSNs, names, and contact data which is 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

   By keeping in mind above information data can be protected

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 faker
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
3745	Timothy	Lee	lindseyclark@example.net	613.679.1219x7353	Male	51.459689	Film/video editor	1.191037	123289.575900	Operations	Spanish
9507	Latasha	George	lauravazquez@example.org	+1-332-257-4064	Female	46.416296	Medical secretary	25.863185	54736.458853	Human Resource	
7319	Tami	Lowery	teresa18@example.org	8549001539	Male	66.765337	Engineer, maintenance (IT)	35.774599	100905.434129	Human Resource	Spanish
5986	Chelsea	Whitehead	christophermiller@example.net	+1-588-755-3235x74234	Female	57.235417	Dietitian	16.586382	88592.109778	Product	German
1560	Alison	Valentine	franklinpamela@example.net	376-722-1045x972	Male	31.062947	Retail manager	9.825018	69849.116763	Operations	
1559	Anna	Holden	barbervicki@example.net	(953)980-7689	Female	27.091926	Teaching laboratory technician	0.227631	89986.087386	Administrative	
580	Gabriel	Tran	bakerfrank@example.com	213.807.6313x1388	Male	68.473171	Regulatory affairs officer	1.594482	176036.492363	Product	
8661	Patrick	Wilcox	olane@example.net	8808388464	Female	31.833390	Private music teacher	3.463169	66310.444362	Marketing	
6011	Kimberly	Parks	richardsonjames@example.org	891-855-9085x628	Male	43.897603	Dietitian	22.730740	197174.239893	Legal	Hindi
7080	Jessica	Bowman	garciasuzanne@example.net	274-412-1461x673	Male	43.733000	Architectural technologist	12.039092	87225.449947	Administrative	
205	Gabrielle	Garcia	roblesjoseph@example.net	550.867.5607x1931	Male	50.683015	Adult nurse	0.000000	70652.708575	Human Resource	
9699	William	Lewis	matthew68@example.com	512-689-0559	Male	40.786542	Therapist, speech and language	10.221172	67189.723611	Human Resource	
8324	Jennifer	Johnson	chasestark@example.com	+1-328-423-2960x80885	Female	23.623940	Training and development officer	3.491676	62256.803110	Human Resource	
2123	Joshua	Johnson	fwall@example.net	990-326-4728x848	Female	27.590351	Market researcher	5.485395	93913.369592	Operations	Spanish
1818	Patrick	Jennings	ichavez@example.com	001-446-580-8979x654	Male	40.608273	Conservation officer, historic buildings	3.154436	151989.535570	I/T	Korean
1834	Marcus	Love	beckyadams@example.com	439-481-0430x1791	Male	36.119211	Advertising copywriter	6.444429	101451.011030	Marketing	
3042	Kevin	Brewer	marvinrhodes@example.org	461.928.3380	Male	41.082202	Senior tax professional/tax inspector	14.899926	109269.481008	Operations	Mandarin, Spanish
5247	Jason	Lopez	stewartalicia@example.org	851.835.5977x9921	Female	38.306228	Engineer, materials	12.858213	65098.058436	Administrative	Tagalog, Mandarin
4319	Matthew	Williams	brooke94@example.com	3513559631	Female	60.244215	Broadcast journalist	34.093479	94154.430791	Marketing	
2912	Michael	Lee	gflores@example.org	772.795.8677x66529	Female	44.350886	Exhibitions officer, museum/gallery	15.595474	56451.369138	Administrative	

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
5390	First Name _	Last	Email_5390	Phone_5390	Male	49	Education officer,	27	119356.21	Operations		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.