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.
Answer:
Publishing detailed ride-share data, even with personal information removed, can still pose significant privacy risks. Here are the main problems:

- If an attacker knows an individual's home and work address, they can identify trips that match these start and end points, revealing commute patterns.
- Trips to and from specific events can be linked to attendees. Publicly available event attendance information can be used for cross-referencing.

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://techcrunch.com/2023/11/16/samsung-hackers-customer-data-breach/#:~:text=Prior%20to%20this%2C%20in%20March,algorithms%20for%20biometric%20unlock%20operations.

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.
Answer:
Summary of Breach:
Samsung experienced a data breach affecting U.K. customers who made purchases on its e-store between July 1, 2019, and June 30, 2020. The breach, caused by a vulnerability in a third-party application, exposed names, phone numbers, postal addresses, and email addresses. It was discovered on November 13, 2023, over three years later. No financial data or passwords were compromised. The incident has been reported to the U.K.'s Information Commissioner's Office (ICO) for investigation. This marks Samsung's third data breach in two years.
Prevention Techniques:
Regular Security Audits:

- Conducting regular security audits and penetration testing can help identify and remediate vulnerabilities in both in-house and third-party applications before attackers can exploit them.

Data Encryption

- Encrypting sensitive data both at rest and in transit can protect information even if an attacker gains access to the systems. This ensures that the data is unreadable without the appropriate decryption keys.

Recommendations:
Enhanced Security Audits:

- Conduct more frequent and comprehensive security audits, focusing on third-party applications to identify and mitigate vulnerabilities promptly.

Data Encryption:

- Ensure all sensitive customer data is encrypted at rest and in transit. This reduces the risk of data exposure even if systems are compromised.

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.

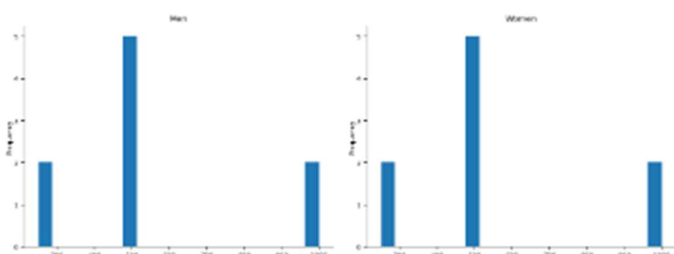
Save your new database to your repository alongside your code that synthesized the data.

C. [SHOULD] Analyze the Synthetic Company

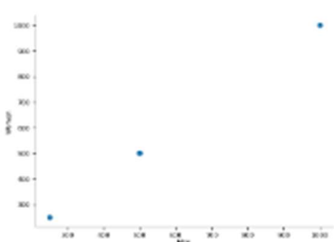
- How many men vs. women will we need to hire in each department?

	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

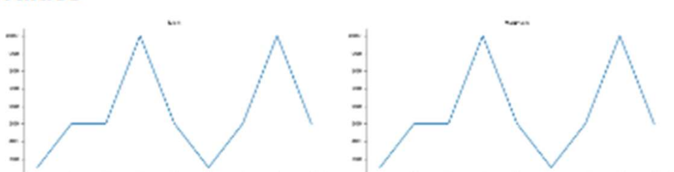
Distributions



2-d distributions



Values



- 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=10000?
  Small businesses should be merged into one entity so that they may hire contract workers or freelancers, thereby increasing the workforce; investments in technology and automation will assist to scale the workforce; remote work can also help to increase it.
- How much office space will this company require?

```
Total office space required: 1,500,000 square feet
```

- Does this new dataset preserve the privacy of the original employees listed in employees.csv?

  The new dataset we've generated potentially compromises the privacy of the original employees listed in employees.csv, especially if it contains real employee data. This dataset includes personal information such as SSNs, names, and contact details, which raises significant data privacy concerns. To ensure privacy, we need to:

  1. Anonymize the data.
  2. Mask sensitive data, known as personally identifiable information (PII).
  3. Provide aggregated data instead of individual records.

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?

- The synthesized data may exhibit a more uniform distribution of games, salaries, and experience levels compared to the original dataset.
- The dataset contains fake email and phone number patterns, indicating a lack of diversity.
- The distribution of languages spoken may not accurately reflect the real data.

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

1. Compare the age, experience, and pay distributions to the original data.
2. Rather of creating it from fictitious data, we may represent true Jon titles from authentic data.

3. Verify that the distribution of the department, age, and gender corresponds with the distribution of the original information

4. Using user phone numbers and emails that are more realistic 5. Using more accurate data for language competence depending on the demographics of the company's workforce

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
5261	Timothy	Williams	amy26@example.net	4052249379	Female	43.691108	Forensic psychologist	16.123203	133916.907238	I/T	Hindi
2822	Brittany	Phillips	mobrien@example.com	+1-301-374-0499x84725	Male	46.163180	Tax adviser	21.857998	138432.453727	Operations	
706	Charles	Livingston	jamestaylor@example.net	+1-316-829-9105x3311	Female	34.955233	Minerals surveyor	12.195512	62005.567616	Sales	Korean
2769	Phyllis	Dillon	daisy45@example.org	814.448.2175	Male	49.061613	Geophysical data processor	8.571820	200373.243469	Product	
5530	Kimberly	Robles	cdavis@example.org	+1-925-435-4004x42778	Female	64.519935	Medical laboratory scientific officer	24.050710	110725.260222	Operations	
2984	Lynn	Henry	lauraburns@example.com	+1-413-370-2914x212	Female	46.135379	Neurosurgeon	14.610998	51454.087944	Sales	Spanish
3691	Lori	Rose	jamesrodriguez@example.org	+1-572-445-5092x63539	Female	56.051264	Chiropractor	7.801127	154114.814715	Product	Korean, Spanish
8542	Christopher	Watkins	jennifer08@example.org	+1-403-308-0018x028	Male	51.229789	Surgeon	25.992859	144480.660011	Marketing	Hindi
8463	Andrew	Smith	hjimenez@example.org	001-582-531-7541x125	Female	60.979694	Agricultural engineer	0.000000	145822.370715	Product	
2749	Samantha	Willis	cynthia51@example.org	751-928-5941	Female	56.178062	Veterinary surgeon	7.106925	153643.285123	I/T	
4610	Thomas	Parks	reynoldssydney@example.net	(886)651-8336x72505	Female	39.031872	Geographical information systems officer	1.095820	123826.719321	Operations	French
2443	Jennifer	Rose	wareleah@example.com	001-317-281-5882	Female	34.861087	Nurse, adult	8.490040	73859.555479	Marketing	
3584	Hayley	Lozano	gonzalezjohn@example.net	769-950-6548	Female	49.039548	Retail banker	0.883948	104062.314796	Product	Hindi
9169	Michael	Terry	svaughn@example.net	648-628-2372	Female	28.296686	Conservation officer, nature	0.551575	136398.398374	Operations	Mandarin
8634	Charles	Lopez	nicolenguyen@example.org	347.968.2957x23871	Male	48.478053	Town planner	9.574344	121021.591260	Marketing	German
1332	Ryan	Rivas	danalambert@example.org	+1-492-640-0558x73003	Male	47.396592	Holiday representative	11.528561	90902.684111	Product	
8056	Austin	Savage	diamondwalker@example.org	(877)636-3648	Female	19.368138	Speech and language therapist	2.331920	119574.717080	Legal	Spanish, German
309	Kristin	Hess	mccarthytasha@example.org	961.977.2252x45244	Female	63.468412	Camera operator	4.889971	180901.163483	Product	
4489	Tina	Allison	jacksonstephanie@example.com	(211)784-8188x2072	Female	51.966763	Clinical molecular geneticist	13.623376	188285.991849	I/T	
2104	Nicholas	Hale	sholmes@example.net	+1-897-627-4307	Male	47.221457	Archaeologist	22.500938	67566.219034	Finance	Mandarin

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: Take into account the normal age range and the amount of variance in the original dataset when selecting the standard deviation for modifying ages. A higher standard deviation might be suitable to add more variance if the dataset spans a wide range of ages. On the other hand, a lower standard deviation ought to be sufficient if the age range is limited. A standard deviation of two to five years is usually a reasonable choice.
2. Salary: Take into account the distribution and variability of wages in your dataset when calculating the standard deviation for modifying salaries. A bigger standard deviation is appropriate if wages are widely distributed and there is a notable disparity in employee compensation. However, a lower standard deviation will do if incomes are generally consistent. Generally speaking, selecting a standard deviation between 5% and 20% of the average salary is reasonable.

Years of Experience: When compensating for years of experience, the standard deviation should represent the dataset's variability. A higher standard deviation is suitable when personnel have a wide variety of experience levels. But if the majority of workers have For comparable years of experience, a lower standard deviation will do. For this reason, a standard deviation of one to three years is usually appropriate.

When perturbing data, you don't have to utilize the same standard deviation for each attribute. Alternatively, you may modify the standard deviation to match the distinct qualities and fluctuations of every attribute in your collection.

You may use a bigger standard deviation for changing age, for instance, if the age range in your data is wider and exhibits more fluctuation than the income range. On the other hand, you would choose a smaller standard deviation for years of experience that are perturbing if they exhibit less variance than age and pay. This method enables more accurate modifications catered to the unique variability of each parameter.

To summarise, the selection of the standard deviation need to be predicated on the attributes present in the dataset, with the objective of including plausible fluctuations while maintaining the general distribution and attributes of the initial data.