DataEng: Data Ethics In-class Assignment

Will McIntosh

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.

- Will -
    - Scenario:
        - Let's say Jane is a public profile
    - Data Collection:
        - Jane's Known Locations: Obtain Jane's home and work address from public records or social media.
        - Ride Data: Get the published ride data with trip details.
    - Pattern Identification:
        - Morning Commute: Identify all trips starting from Jane's home to her work address between 7 AM and 9 AM.
        - Evening Commute: Identify trips from work to home between 5 PM and 7 PM.
    - Correlate Vehicle Data:
        - Car Match: Narrow down trips where the vehicle matches the one Jane was seen in on social media or public events.

    - Final Match:
        - Unique Routes: Confirm trips by matching unique routes Jane might take that are not common among other riders.
        - Consistency Check: Ensure the identified trips form a consistent pattern over weeks or months, reinforcing the likelihood they belong to Jane.
    - Interest:
        - A stalker might pay for information on a celebrity's rides to follow them or figure out their daily routine, posing a significant threat to their safety.
- John
    - The problem with this approach is that it's easy to correlate the data to find private information. For example, if someone wants to stalk someone from a grocery store or gym, and they notice this person uses the rideshare company, all they need to do is get the start locations of the rides that end at the gym or grocery store, then they can see where the person lives or works.
- Chase
    - 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 like homes.
- Trae
    - The most dangerous portion of the information that was given to the general public is the start/end location. If it comes to the more generalized places like grocery stores or public venues then it wouldn't matter as much, but when it comes to more residential areas, that information can be cross referenced fairly easily. If we include the timestamp and regularity of the ride there is a chance that a person's schedule could be extrapolated based on that data.
    - Overall regardless of the efforts that have been made in order to anonymously release this information, it is too easily referenced for other purposes. Even if it is someone coming from a public place and going somewhere else that could be easily trackable with data like that. You could not only track their comings and goings from that place, but assume based on timing the other trips they make via this riding service.

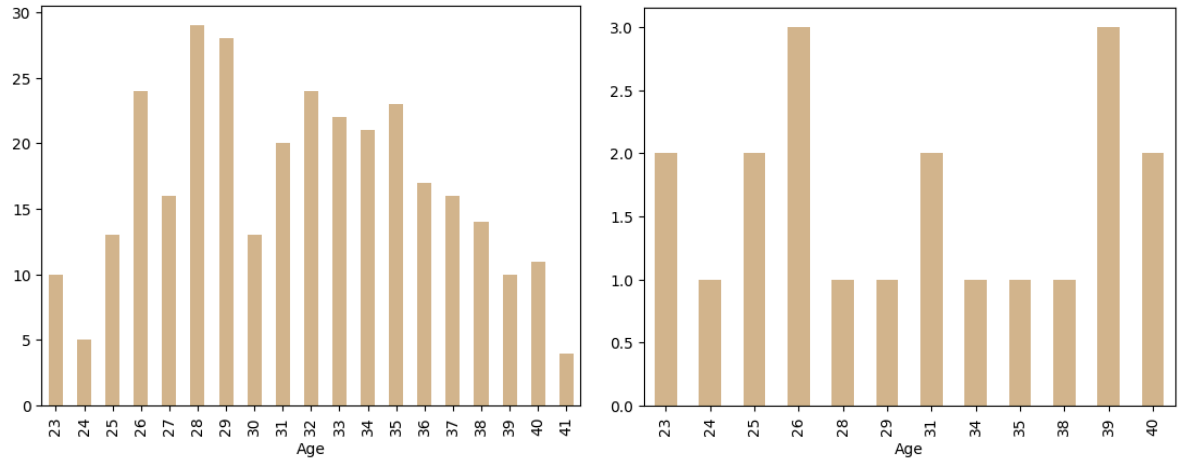
Search the internet and provide a URL of one article that describes one data breach that occurred during the previous 5 years. The breach must be one in which the attacker obtained personal, private information about customers or employees of the attacked enterprise.

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.

- **Data Minimization:** Some of the data that Capital One kept on file that they didn't need was:
    - **Credit Card Application Data:** Capital One retained information from credit card applications, including names, addresses, zip codes, phone numbers, email addresses, dates of birth, and self-reported income.
    - **Social Security Numbers and Bank Account Numbers:** Approximately 140,000 Social Security numbers and 80,000 linked bank account numbers were exposed.
    - **Customer Status Data:** Information such as credit scores, credit limits, balances, payment history, and contact information was also accessed.
    - **Transaction Data Fragments:** The attacker obtained fragments of transaction data from several years.

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.

	First Name	Last Name	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department
114	4393625390128698851	4387725337542742351	10238666586886478251	-3066532837100290344	male	39	Machine Learning Engineer	13	11500	Product
306	-8442619924142601169	-1394951485486561113	60739950575710698 51	71203131508997115 70	female	26	Web Developer	2	8000	Product
229	-5579788096918675528	2792649405253931294	85383684325501189 71	21082393838611214 6	male	25	Mobile Developer	1	7500	Product
185	-4893865832598899061	3771786640846578929	32539916657513250 92	-5196901923736111712	male	28	Web Developer	2	8000	Product
52	2063459458923908530	-4972007852729567929	-7233130027272198191	5088719323130514951	male	39	Designer	14	15000	Product
16	674497897022979647	7235858257978037538	-1552426665629328232	-5622140270837563728	female	26	Designer	4	10000	Product
282	3452751507884330355	3319281907048874850	-5215276553332020852	-475533335173405170	male	26	Web Developer	2	8000	Product
186	-4893865832598899061	2845741082281447465	-2537958055310679700	963645498250851279	male	29	Mobile Developer	5	7500	Product
222	-4893865832598899061	451590478997412045	-6749959797873042347	8813938745029318187	male	38	Mobile Developer	12	12000	Product
8	-285127793273212711	11056468565492002 15	90689567909894393 22	57558522276812832 45	female	34	Web Developer	10	11000	Product

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?

	First Name	Last Name	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department
114	4393625390128698851	4387725337542742351	10238666586884782 51	-3066532837100290344	male	39.320327	Machine Learning Engineer	11.304863	7951.939010	Product
306	-8442619924142601169	-1394951485486561113	60739950575710698 51	71203131508997115 70	female	25.448224	Web Developer	1.254000	9513.502621	Product
229	-5579788096918675528	2792649405253931294	85383684325501189 71	21082393838611214 6	male	25.380452	Mobile Developer	1.130595	8186.401066	Product
185	-4893865832598899061	3771786640846578929	32539916657513250 92	-5196901923736111712	male	27.848261	Web Developer	3.702918	17592.497533	Product
52	2063459458923908530	-4972007852729567929	-7233130027272198191	5088719323130514951	male	39.811197	Designer	13.338846	15748.901868	Product
16	674497897022979647	7235858257978037538	-1552426665629328232	-5622140270837563728	female	24.618595	Designer	5.344836	4591.708624	Product
282	3452751507884330355	3319281907048874850	-5215276553332020852	-475533335173405170	male	25.553046	Web Developer	0.995535	-65.093671	Product
186	-4893865832598899061	2845741082281447465	-2537958055310679700	963645498250851279	male	30.675133	Mobile Developer	4.720650	19135.424732	Product
222	-4893865832598899061	451590478997412045	-6749959797873042347	8813938745029318187	male	38.385262	Mobile Developer	12.434184	11942.773297	Product
8	-285127793273212711	11056468565492002 15	90689567909894393 22	57558522276812832 45	female	33.260287	Web Developer	10.531267	7193.625795	Product

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~~

- [x] ~~Need to keep track of social security numbers~~
- [x] ~~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.~~
- [x] ~~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: https://www.uscis.gov/sites/default/files/document/data/h-1b-petitions-by-gender-country-of-birth-fy2019.pdf~~
- [x] ~~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.~~
- [x] ~~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~~
- [x] ~~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.

```
expanded_df.shape=(10000, 13)
```

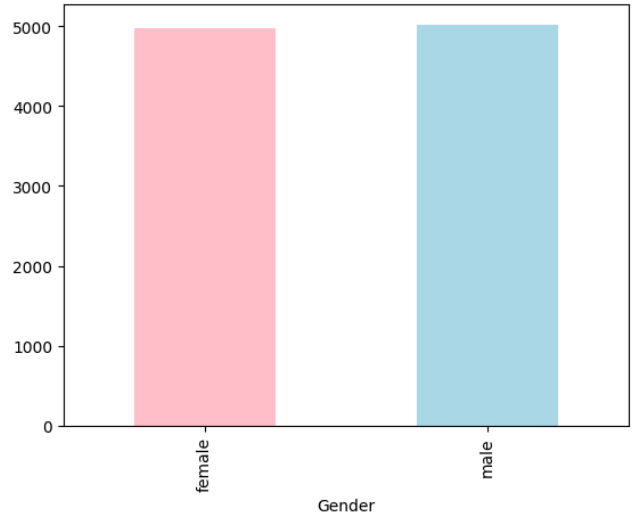
	First Name	Last Name	Email	Phone	Gender	Age	Job Title	Years Of Experience	Salary	Department	SSN	Languages	Country
0	Jose	Lopez	joselopez0944@slingacademy.com	+1-971-533-4552x1542	male	25	Project Manager	1	8500	Product	354-03-4009		Taiwan
1	Diane	Carter	dianecarter1228@slingacademy.com	881.633.0107	female	26	Machine Learning Engineer	2	7000	Product	424-06-8862	French	India
2	Shawn	Foster	shawnfoster2695@slingacademy.com	001-966-861-0065x493	male	37	Project Manager	14	17000	Product	427-04-6986	Tagalog	Taiwan
3	Brenda	Fisher	brendafisher3185@slingacademy.com	001-574-564-4648	female	31	Web Developer	8	10000	Product	542-36-2718	Korean, German	India
4	Sean	Hunter	seanhunter4753@slingacademy.com	5838355842	male	35	Project Manager	11	14500	Product	661-05-7525		India

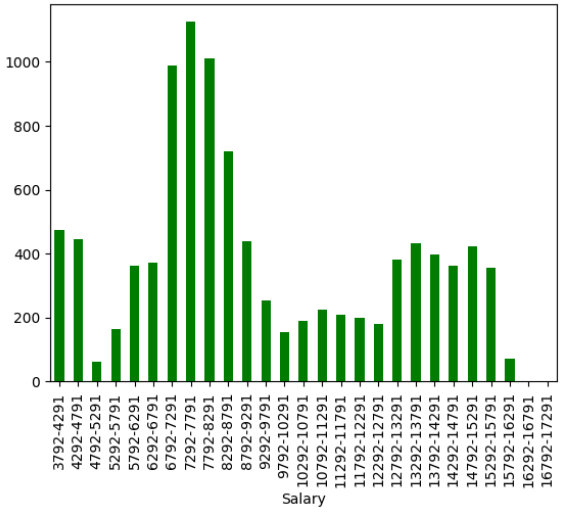
G. [SHOULD] Analyze the Synthetic Company

- 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?

Department	Current Number of Men	Current Number of Women	Current Male Ratio	Current Female Ratio	Men Hiring Need	Women Hiring Need	Future Number of Men	Future Number of Women	Future Male Ratio	Future Female Ratio
Administrative	478	479	0.499478	0.500522	1.0	0.0	479.0	479.0	0.5	0.5
Finance	262	244	0.517787	0.482213	0.0	18.0	262.0	262.0	0.5	0.5
Human Resource	511	485	0.513052	0.486948	0.0	26.0	511.0	511.0	0.5	0.5
I/T	456	448	0.504425	0.495575	0.0	8.0	456.0	456.0	0.5	0.5
Legal	214	259	0.452431	0.547569	45.0	0.0	259.0	259.0	0.5	0.5
Marketing	482	473	0.504712	0.495288	0.0	9.0	482.0	482.0	0.5	0.5
Operations	978	979	0.499745	0.500255	1.0	0.0	979.0	979.0	0.5	0.5
Product	1141	1128	0.502865	0.497135	0.0	13.0	1141.0	1141.0	0.5	0.5
Sales	501	482	0.509664	0.490336	0.0	19.0	501.0	501.0	0.5	0.5





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?