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.

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.

### **Potential Problems with Publishing Ride Data**

* **Location Data**: Start and end locations can reveal home or workplace addresses.
* **Temporal Data**: Time stamps combined with location data can identify routines.
* **Vehicle Information**: Make, model, and year can narrow down drivers, indirectly identifying passengers.
* **GPS Breadcrumbs**: Detailed routes can be matched with social media or public records to identify individuals.

### **How an Attacker Might Re-identify Passengers**

* **Cross-referencing**: Matching ride data with social media or public records.
* **Frequent Routes**: Identifying patterns like home-to-work trips.
* **Unique Locations**: Trips to specific, unique places can identify individuals.

### **Recent Data Breach Example**

* **Breach**: Marriott International Data Breach (2020)
* <https://techcrunch.com/2020/03/31/marriott-hotels-breached-again/>
* **Summary**: Attackers accessed data of 5.2 million guests using compromised employee login credentials, obtaining contact details and personal information.

### **Prevention Techniques**

* **Data Anonymization**: Apply k-anonymity, l-diversity, or t-closeness.
* **Differential Privacy**: Add noise to data to protect individual privacy.
* **Aggregation and Generalization**: Use broader geographical and temporal data.
* **Access Controls and Monitoring**: Restrict and monitor data access.
* **Data Minimization**: Only publish essential data for research purposes.

## B. [MUST] Model Based Synthesis

Your job is to synthesize a data set based on [the employees.csv data set](https://drive.google.com/file/d/1Ki1um7a22Gw2iXC37q4FcnB8qL4pQW_k)

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](https://www.uscis.gov/sites/default/files/document/data/h-1b-petitions-by-gender-country-of-birth-fy2019.pdf).
* 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?

Female:

Sales: 499

Operations: 1043

Marketing: 505

Product: 970

Administrative: 499

Human Resource: 474

Finance: 274

I/T: 513

Legal: 223

Male:

Sales: 489

Product: 1022

Operations: 971

Legal: 247

I/T: 521

Human Resource: 504

Administrative: 515

Marketing: 480

Finance: 251

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

$945932154

* Other than hiring from non-US countries, how else might the company grow quickly from size=320 to size=10000?

Focusing on organic growth by expanding into new markets, launching new products or services, and increasing market share can lead to significant expansion over time.

Investing in technology and digital transformation can improve efficiency and productivity, allowing the company to scale more quickly.

* How much office space will this company require?

1000000 sq ft

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

0

## 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 dataset comprises only text, categorical, and numeric variables. In a real company dataset, we would expect a wider variety of variable types, including datetime, boolean, and more specific categorical types. The correlations between variables in this synthetic dataset are likely too perfect or controlled, unlike the more nuanced and varied correlations found in real datasets. Furthermore, the synthetic dataset shows a high imbalance in the department variable (79.9%), which appears too uniform compared to the typically more diverse and naturally distributed departments in a real company.

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

We can introduce noise or randomness to the data to make it more diverse and less uniform. For instance, slight variations can be added to salaries or years of experience. Instead of using highly controlled correlations, we can employ statistical techniques to create correlations that better reflect real-world relationships between variables. Additionally, we can broaden the range of variable types in the dataset to include datetime, boolean, and other common types found in real-world datasets. Introducing missing data can simulate the incomplete nature of real datasets, making the dataset more realistic and allowing for the testing of missing data handling techniques.

## 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.

## F. [SHOULD] Anonymization

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

## 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?

When adding Gaussian noise to the age, salary, and years of experience attributes in the employee data, the standard deviation parameter should be selected based on the desired degree of perturbation. It's not essential to use the same standard deviation for all three attributes. Instead, the standard deviation should reflect the scale and variability of each attribute. For instance, since salary values typically exhibit greater scale and variability than age values, a larger standard deviation can be applied to salary than to age.