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.

John - The problem with this approach is that if a stalker has other information on the riders, they can easily correlate the data to find out the whereabouts and routine of someone. 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.

```
### 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.
```

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://time.com/6984811/ticketmaster-data-breach-customers-livenation-everything-to-know/

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.

This was a data breach that happened to ticketmaster. This actually happened very recently. The breach was reportedly done by a group called ShinyHunters who state it is seeking $500,000 for the 1.3TB database of compromised customer data, which it claims includes names, addresses, phone numbers, and credit card details of 560 million users.

There could have been multiple things done to prevent this. The first is encrypting the data more efficiently. The second could have been performing tests on the databases to see how easily they were penetrated then again, fixing those issues. I think maybe some of the data could have dropped from the database. Yes, the billing information is needed for transactions but specifically we don’t need to store it, unless users don’t want to put in their information every time.

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.

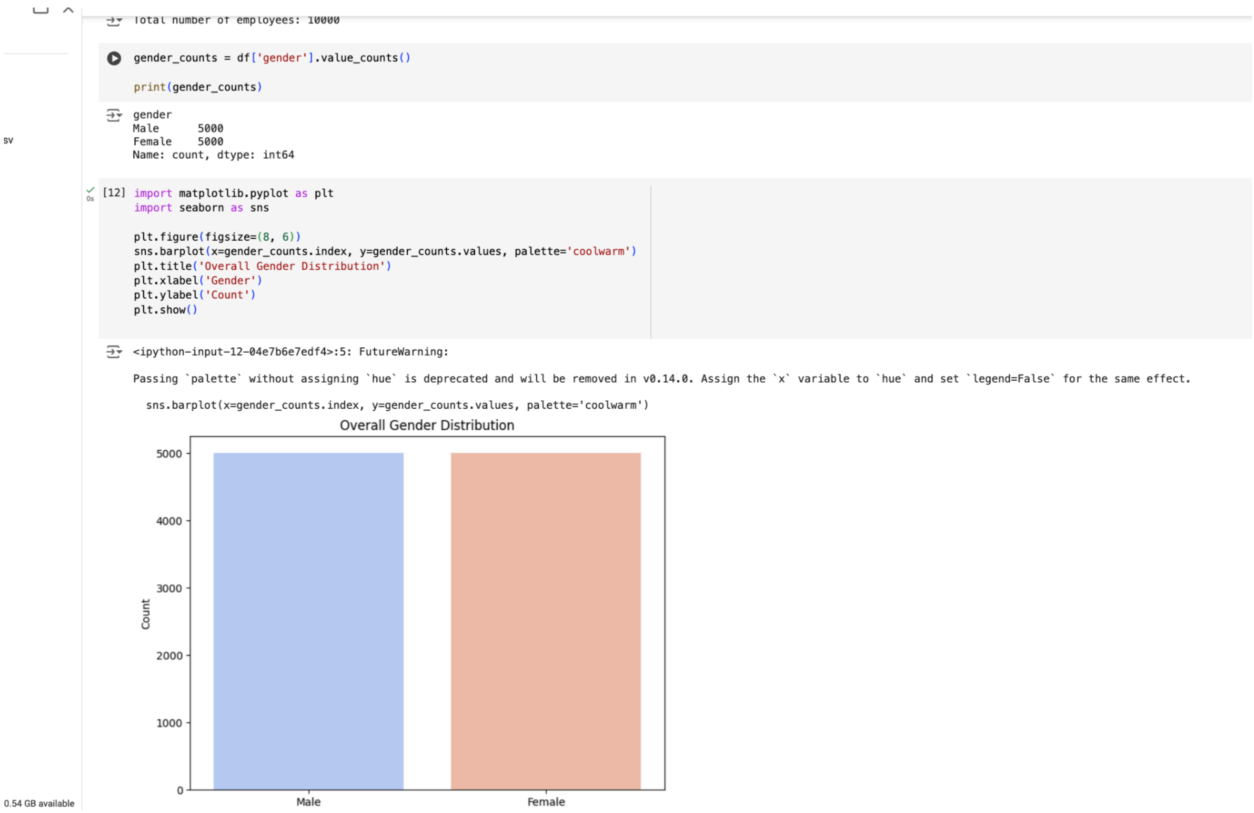
Use the Faker python module to produce a 10K employee dataset. Follow these constraints:

- 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 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 company has exactly 5000, male and female employees



```
Total number of employees: 10000
```

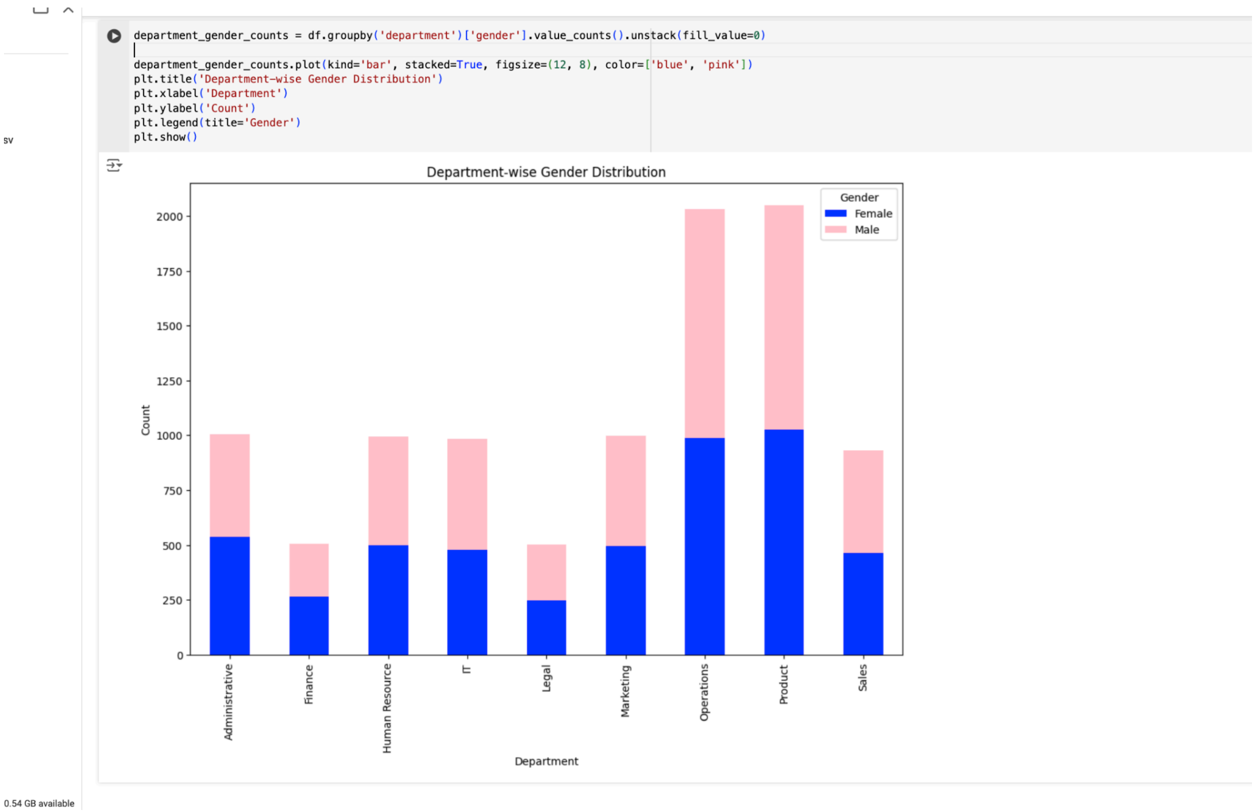
```
gender_counts = df['gender'].value_counts()

print(gender_counts)
```

```
gender
Male      5000
Female    5000
Name: count, dtype: int64
```

```
[12] import matplotlib.pyplot as plt
     import seaborn as sns

     plt.figure(figsize=(8, 6))
     sns.barplot(x=gender_counts.index, y=gender_counts.values, palette='coolwarm')
     plt.title('Overall Gender Distribution')
     plt.xlabel('Gender')
     plt.ylabel('Count')
     plt.show()
```

```
<ipython-input-12-04e7b6e7edf4>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=gender_counts.index, y=gender_counts.values, palette='coolwarm')
```

For the whole company now per department it would be:



```
department_gender_counts = df.groupby('department')['gender'].value_counts().unstack(fill_value=0)

department_gender_counts.plot(kind='bar', stacked=True, figsize=(12, 8), color=['blue', 'pink'])
plt.title('Department-wise Gender Distribution')
plt.xlabel('Department')
plt.ylabel('Count')
plt.legend(title='Gender')
plt.show()
```

Looks approximately the same, but I do see a slight difference in the administrative department.

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

```
Total Payroll: $829,374,600.00
```

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

Not sure how to relate this question to the data, but I think that if the company would decide to go harder on recruiting, as well as offer bonus incentives, that would allow the company to prosper. I think adding new departments (thus adding new roles would also help).

- How much office space will this company require?

“According to Office Finder, it's best to allow about 175 square feet of space for each person who works in your office” - https://thereceptionist.com/blog/how-much-office-space-do-you-really-need/#:~:text=According%20to%20Office%20Finder%2C%20it%27s,Admittedly%2C%20that%27s%20tough%20to%20visualize.

Assuming this if we have 10,000 employees, the office space would be 175*10,0000. However not all of these will work on site or in office, so a safe assumption would be to say 5,000 employees do. This would mean that the company would require at least 5,000*175 = 875,000 square feet of office space. This is assuming only have of the workers work in office so >875,000

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

Yes it does because none of the employees from the original dataset are in there.

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?

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

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?