DataEng: Data Ethics In-class Assignment

A. Discussion Questions

Can you see a problem with this approach? How might an attacker re-identify some of the real passengers?

Response:

Although all the personal information is stripped from the data before sharing the data still contains the start and end location and the gps bread crumb data. It is possible that a person may have reserved a trip multiple times from the same start location or may travel to the same end location. Based upon the start location and end location it would be easy to identify the passenger by integrating the data from other sources that may have some identifiable information or sometimes we can identify people from the other columns.By correlating these sources, an attacker could uncover the identities of passengers, thus compromising their privacy.

Data Breach: *https://www.helpnetsecurity.com/2023/05/03/t-mobile-breach-2023/*
Description: T-Mobile experienced a cyber attack from February 24 to March 30, 2023, affecting 836 customers. The company's security measures detected unauthorized access, as some bad actor obtained some information from a small number of accounts. The compromised data varied for each customer and may have included personal information like full name, contact details, account numbers, T-Mobile account PIN, social security number, government ID, date of birth, balance due, and internal codes used by T-Mobile for customer accounts. However, no personal financial account information or call records were compromised.

To prevent data breaches, it's important to encrypt sensitive customer data. This ensures that even if data is intercepted or accessed by hackers, it remains protected and unreadable. Implementing strict access control measures is also crucial to limit who can access sensitive information, reducing the risk of unauthorized access. Regular security audits and vulnerability assessments should be conducted to identify and fix potential security weaknesses before attackers can exploit them. Providing developers with training and tools to write secure code helps minimize vulnerabilities in software. Also, having robust data backup and disaster recovery plans ensures that data can be restored quickly in the event of a breach.

C. Analyze the Synthetic Company

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

```
Men vs. Women in each department:
Gender          female  male
Department
Administrative     509   508
Finance            258   227
Human Resource     472   514
I/T                490   501
Legal              255   229
Marketing          499   512
Operations        1000  1018
Product            988  1006
Sales              487   527
```

  Based on the count of employees from each department we can decide on how many employees to hire in each department.

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

```
Yearly payroll:
89268054
```

- **Other than hiring from non-US countries, how else might the company grow quickly from size=320 to size=10000?**
  The company can increase its employee count by hiring more employees based on the ratio of male to female employees.Also, it can increase the number of employees in each department based on the current count of employees in those departments.

- **How much office space will this company require?**

```
Total office space required (square feet):
200000
```

- **Does this new dataset preserve the privacy of the original employees listed in employees.csv?**
  As we can observe that there are no overlapping records the privacy of the employees listed in the employees.csv file is preserved.

```
Number of overlapping records between datasets:
0
```

D. Quality of the Synthetic Dataset

In what ways does the synthetic data set appear to be obviously synthetic and/or not representative of the current company?

Response:

The mean salary in the original dataset is 10175, whereas the mean salary in the synthetic dataset is significantly lower at 8921.65. This discrepancy suggests that the salary distribution in the synthetic data is not aligned with that of the original dataset.The mean age of employees in the original dataset is 31.59 years, while in the synthetic dataset, it is substantially higher at 43.36 years.This indicates that the age in the synthetic dataset may not accurately reflect the company's workforce demographics.The original dataset contains employees from 8 distinct departments. In contrast, the synthetic dataset includes 638 distinct job titles, which suggests a lack of proper alignment with the original department structure.

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

Response:

We can modify the data generation process to reflect the age distribution of the original dataset. This can be done by either sampling directly from the original data or by analyzing the statistical patterns and using them to generate new, representative data.

G. Perturbation

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?

Response:

We can select a small standard deviation value for all attributes to ensure that the original dataset distribution is preserved. However, the standard deviation values do not need to be identical for each attribute. For instance, we could choose a small deviation that reflects a 2-3 year difference for attributes like age or years of experience.