CS211 - Data Privacy: Final Project

Vincent Moeykens

The University of Vermont

Fall 2020

Project Goals

This project had a few simple goals:

1. Identify a data with potential PII
2. Determine which statistics would be valuable from this data
3. Create a sysytem to generate differentially private statistics from this data (using a reasonable privacy budget: ϵ)
4. Automatically generate a PDF report when finished

Dataset

The dataset I used from this project comes from *Kaggle*. The dataset contains information about credit card customers at some company. While the data itself does not contain simple PII such as names, dates of birth, or SSN; it does contain information such as:

- Customer Age

Dataset

The dataset I used from this project comes from *Kaggle*. The dataset contains information about credit card customers at some company. While the data itself does not contain simple PII such as names, dates of birth, or SSN; it does contain information such as:

- Customer Age
- Customer Gender

Dataset

The dataset I used from this project comes from *Kaggle*. The dataset contains information about credit card customers at some company. While the data itself does not contain simple PII such as names, dates of birth, or SSN; it does contain information such as:

- Customer Age
- Customer Gender
- Education Level

Dataset

The dataset I used from this project comes from *Kaggle*. The dataset contains information about credit card customers at some company. While the data itself does not contain simple PII such as names, dates of birth, or SSN; it does contain information such as:

- Customer Age
- Customer Gender
- Education Level
- Marital Status

Dataset

The dataset I used from this project comes from *Kaggle*. The dataset contains information about credit card customers at some company. While the data itself does not contain simple PII such as names, dates of birth, or SSN; it does contain information such as:

- Customer Age
- Customer Gender
- Education Level
- Marital Status
- Income Level

Dataset

The dataset I used from this project comes from *Kaggle*. The dataset contains information about credit card customers at some company. While the data itself does not contain simple PII such as names, dates of birth, or SSN; it does contain information such as:

- Customer Age
- Customer Gender
- Education Level
- Marital Status
- Income Level
- **All things that could be used to re-identify individuals!**

Statistics

I picked a few basic statistics that I thought an analyst might be interested in. They are:

- Average Customer Age
- Average Months on Book
- Average Credit Limit
- Count of most common Income Ranges
- Count of most common Education Level
- Average Credit Limit of Customers younger than 33y/o vs Average Credit Limit of Customers older than 33y/o
- Most Common Income Range of Customers with a College Degree vs Customers Without

Privacy Strategy

To generate the differentially private statistics, I used a few different strategies.

For all average statistics, I first use the *above_threshold* method to deteremine a upper clipping parameter. I then generate noisy sums and counts using the laplace mechanism (and a portion of the total epsilon alloted for this query). I then divide the noisy sum by the noisy count to get a differentially private average by post processing.

To determine the category that has the maximum number of occurrences for a parameter, I use the report-noisy-max method.

Report Generation

Finally, I calculate error percentages for average queries, and verify that the report noisy max results match the expected results. This data then gets automatically generated into a .tex file, and compiled to a pdf (if the user has a LaTeX compiler installed)