Exploration of Differential Privacy on Brazilian Accident Data

Alyssa Ahn

May 3, 2022

1 Introduction

In this project, I worked with Brazilian Accident Data from the Ministry of Justice and Public Security. This data set contains accidents in which the police responded to accidents that occurred on Brazilian Federal Highways. While this reporting provides accident data, the aggregate of occurrences of accidents in any span of time may not be comprehensive. For example, the data set may not include unreported accidents. The data set has been de-identified, meaning that all Personal Identifiable Information (PII) has been scrubbed. In this exercise, I will be performing a review of working definitions of differential privacy, applying a differentially private mechanism to a data set, and analyzing the effects of said mechanism.

Note: 'Mechanism' will be used in definitions and throughout this write up. This term is synonymous to algorithm.

2 Background

2.1 Qualitative Working Definition

Differential Privacy is a measure. A Differential Private Mechanism is able to provide accurate information about a data set while also ensuring a high level of privacy.

2.2 Quantitative Working Definition

Definition 1 *Differential Privacy*
A randomized function κ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subset Range(\kappa)$,

$$Pr[\kappa(D_1) \in S] \leq exp(\epsilon) \times Pr[\kappa(D_2) \in S]$$

2.3 Understanding Epsilon

The definition above can be re-written as:

$$\frac{Pr[\kappa(D_1) \in S]}{Pr[\kappa(D_2) \in S]} \leq exp(\epsilon) \text{ [1]}$$

In rewriting the definition, we observe that ϵ can be expressed as a bounded ratio.

But, what exactly is ϵ measuring and are there any constraints on what it can be or cannot be? ϵ measures the loss of privacy. In fact, ϵ is a publicly known value and we can receive some guidance as noted below.

According to Cynthia Dwork, the Founding Mother of Differential Privacy: "The choice of ϵ is essentially a social question...we tend to think of ϵ as, say, 0.01, 0.1, or in some cases, $ln(2)$ or $ln(3)$" (3). [2] There are many published articles that provide guidance on how to pick ϵ, but the easiest way to think about it is the following inverse relation: Taking ϵ to be small like 0.01 upholds more privacy for the data set in question by adding more noise. While taking ϵ to be large like $ln(2)$ will reveal more about the data set in question in turn and add less noise. We will follow up on this point in the coming sections.

Along with ϵ, we must also consider *sensitivity*.

Definition 2 *Sensitivity*
For $f : \mathcal{D} \mapsto R^k$, the sensitivity of f is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \text{ for all } D_1, D_2 \text{ differing in at most one element.}$$

As noted in Definition 2, we observe that Δf is 1. We observe here that sensitivity is not based off the the cardinality or size of the data set that we are using. In fact, we can say that sensitivity is independent of the data set.

In this project, we worked with the query: "how many accidents occurred in 2019 in each of the Brazilian states." In this case, the metric we are looking at is an aggregate of all the occurrences based off the state code column within the data set. Given that this set does not contain PII, we will create a mechanism that will calculate the various aggregates and focus on protecting the 'privacy' of a single accident.

3 Analysis

3.1 Noise

We have discussed the query at hand-which is looking at the counts of accidents in the various Brazilian states. In order to further uphold anonymity of the data set, we need to explore the topic of noise. With the ϵ definition of Differential Privacy, we are going to leverage the Laplace Mechanism, which is commonly used for histogram queries. When we look at the outputs of the different aggregates, we can use the following equation to capture the impact of the privacy mechanism: $f(X) + Lap(\frac{\Delta f}{\epsilon})$

3.2 Bar Chart Result

Leveraging the code in my GitHub Repository, I set $\epsilon = 0.01$. In doing so, we are ensuring privacy of the data set while also adding more noise. Here is the bar chart produced displaying the two aggregates - the actual count and the count using the differentially private mechanism.

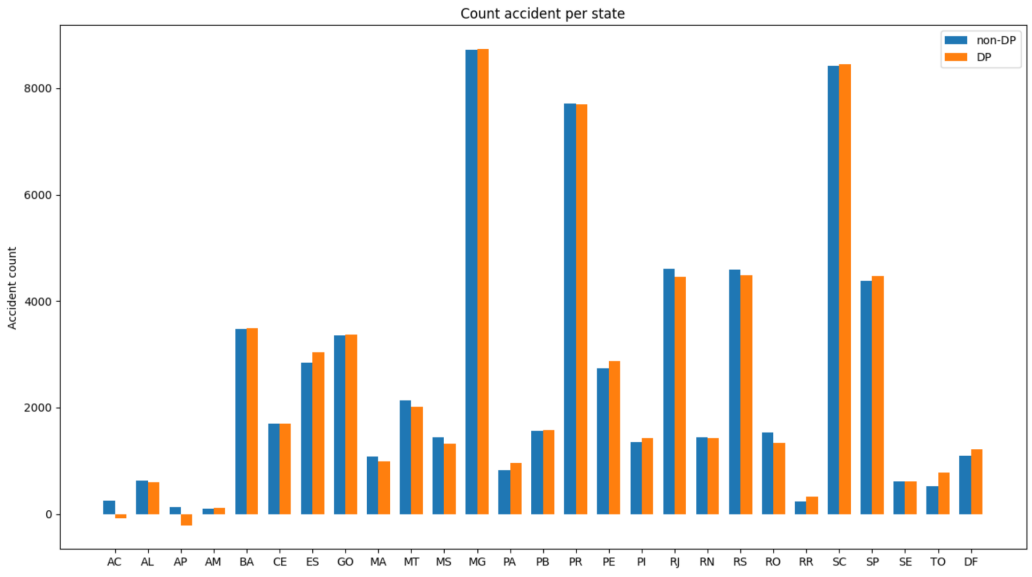

Figure 1: Distribution of Accidents in 2019 by Brazilian State ($\epsilon = 0.01$)

We observe that the counts are relatively close to one another, but there are a few items that may look strange. For example, Amapá (AP) has a count for the number of accidents that occurred in 2019. When we ran the mechanism, we observe that the noise ended up completely removing the count

and gave us a negative count. Given that the sampling from the Laplace distribution is unbiased and random, there is always the possibility that the count for a state may return negative.

Along with the bar charts, I used a map of Brazil to provide shading indicating greater occurrences of accidents. Below is the spatial data analysis charts for actual vs. the differentially private aggregates. As we see, the noise added does not obfuscate the data too far from its true value. Most notably, from the perspective of spatial data, the results of the mechanism appears to be similar to the actual counts per state.

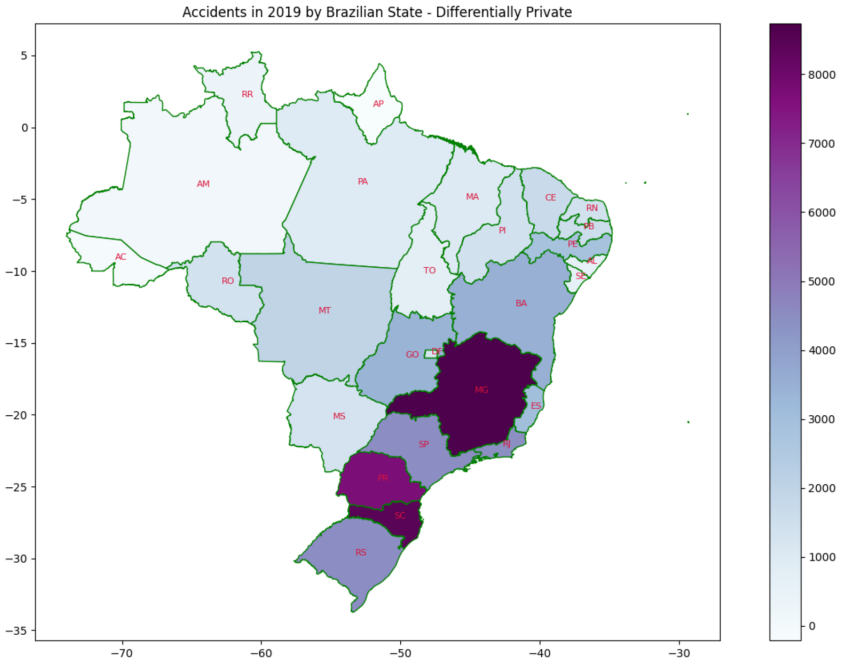

Figure 2: Spatial View: Distribution of Accidents in 2019 by Brazilian State ($\epsilon = 0.01$)

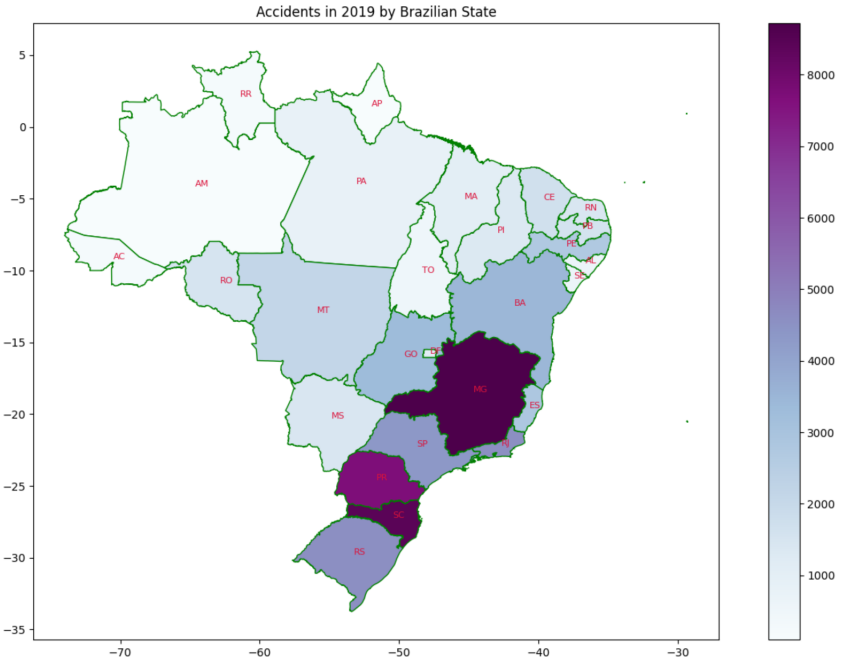

Figure 3: Spatial View: Actual Distribution of Accidents in 2019 by Brazilian State

One other interesting plot that I wanted to study is the Relative Error vs. the Actual Count. This study allows for us to have a better understanding of how the size of the aggregate impacts noise. From 3.2, we observe that states with smaller counts often more strongly affected by the noise, resulting in a larger relative error in comparison to that of other states with larger accident counts.

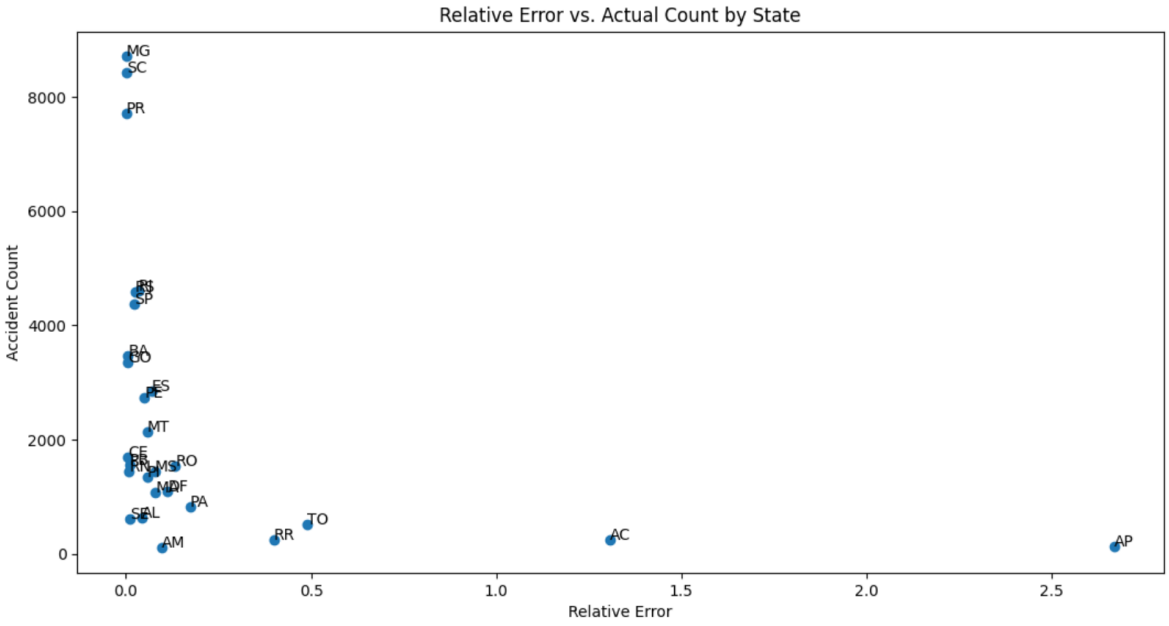

Figure 4: Plot of Relative Error vs. Actual Count ($\epsilon = 0.01$)

3.3 Playing with ϵ

In Section 2.3 we discussed the importance of ϵ and how it impacts both sensitivity and the noise added by the Laplace Mechanism. Let's see what happens when we take $\epsilon = 1$ and explore how it impacts relative error.

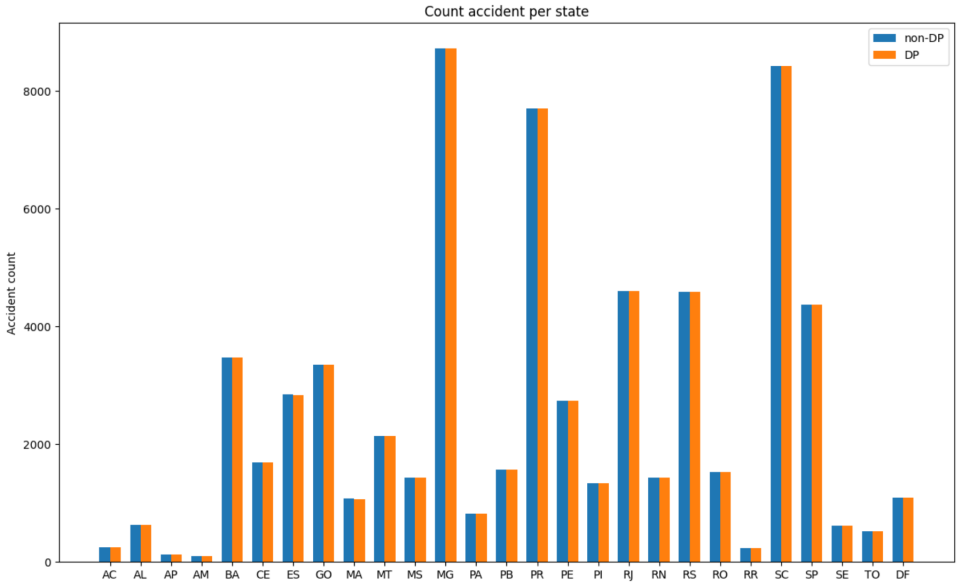

Figure 5: Distribution of Accidents in 2019 by Brazilian State ($\epsilon = 1$)

In a similar fashion as when $\epsilon = 0.01$, the spatial data analysis shows us that the mechanism returns counts that are similar to the actual counts per state. Without any prior knowledge, one would assume that these two maps are indistinguishable!

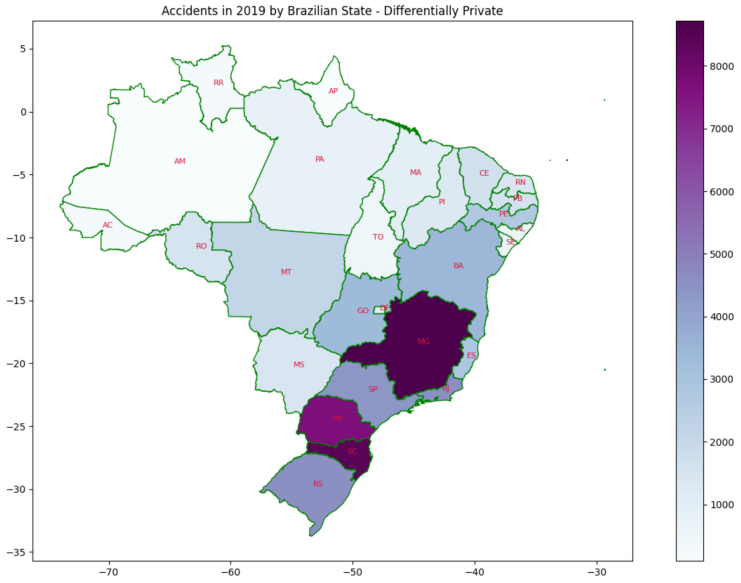

Figure 6: Spatial View: Distribution of Accidents in 2019 by Brazilian State ($\epsilon = 1$)

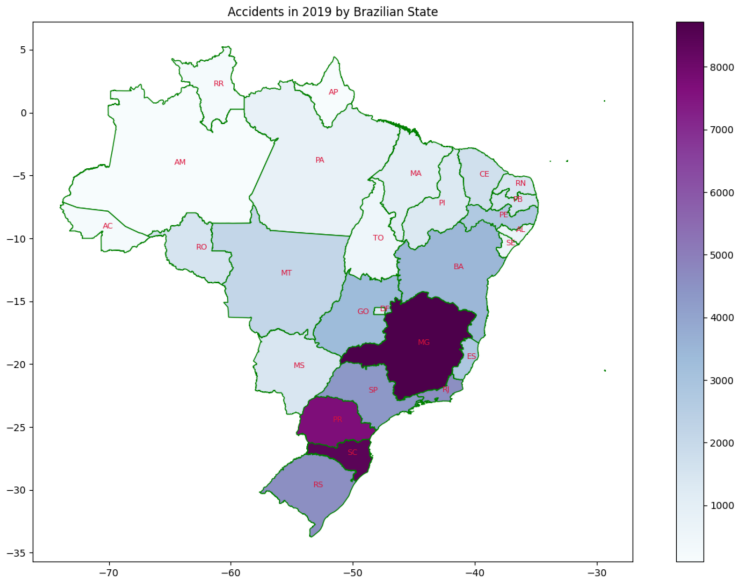

Figure 7: Spatial View: Actual Distribution of Accidents in 2019 by Brazilian State

3.4 Error Analysis for $\epsilon = 1$

We observe once again that the overall behavior of this scatter plot reveals that the "line of best fit" would be the Multiplicative Inverse, or as written $f(x) = \frac{1}{x}$. We also observe that the x-axis operates on a smaller scaling. In the earlier case, we are working with scale factors of 0.5, and now we have a scale of 0.005. With such small relative error, we observe that less noise is applied, which intuitively shows that larger ϵ implies weaker privacy. The largest error observed is with Amapá with an error of 0.02. With such a low error, one must observe that picking $\epsilon = 1$ demonstrates the inability to uphold the same level of privacy as taking ϵ to be sufficiently small.

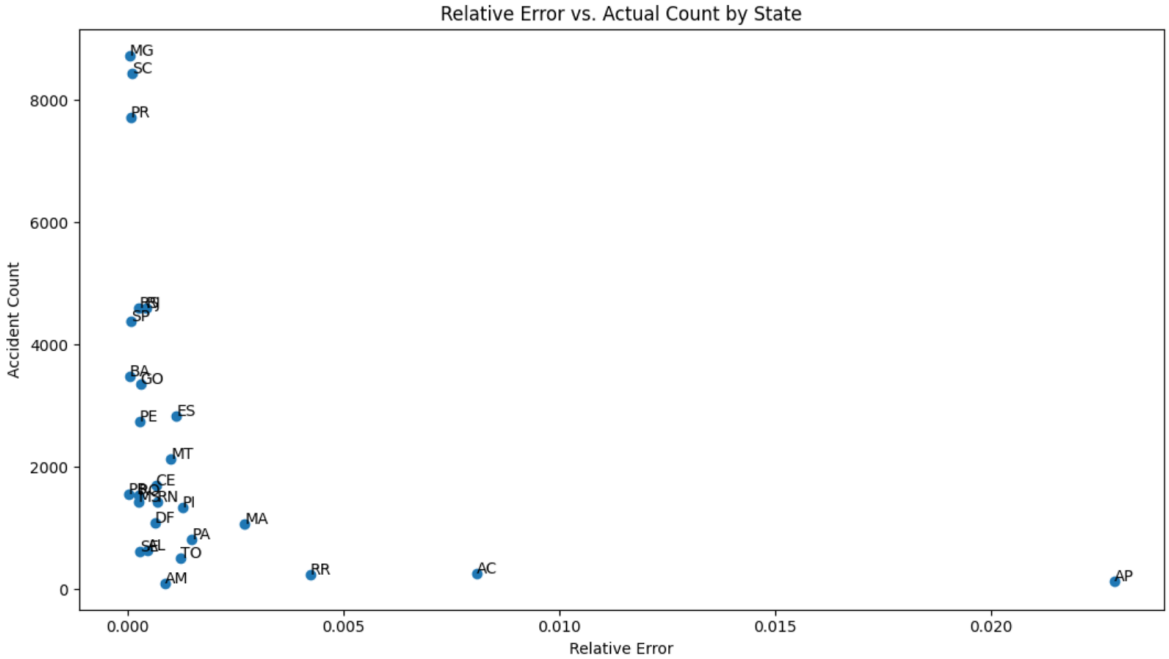

Figure 8: Relative Error vs. Actual Count ($\epsilon = 1$)

4 Conclusion

This was a great project to dabble with Differential Privacy. In this project, I was able to read relevant papers within the field while also being able to apply this mechanism to a data set. I would like to continue to read relevant articles around Differential Privacy and think about how to better select ϵ. Additionally, I would like to ascertain a better understanding of when to use certain mechanisms. I will be exploring δ and its impact on privacy preserving mechanisms. While we use the Laplace mechanism in this project to create noise, I would like to explore the value of leveraging a Gaussian Distribution for noise. In my code, we observe that the metric is count. I would like to think of other types of metrics that can benefit from privacy preservation.

As for future explorations with the current data set, I would want to look at making the data set sparse by privatizing accidents based off of the mile marker column. In that exploration, I do not think that the ϵ Differential Privacy definition would be the most effective manner to anonymize the data. I think that picking the partitions would be less intuitive and I would be concerned about clustering around certain mile markers. One final interest of further exploration would be combining my understanding of the K-Means problem and leveraging the clustering along with privacy mechanisms.

5 Appendix

Here lies the charts that support the relative errors performed with the different ϵ values. This section is an offering of how much error was created due to the noise that was statistically calculated.

State	Relative Error
AC	1.3076
AL	0.0432
AP	2.6684
AM	0.0957
BA	0.0059
CE	0.0054
ES	0.0692
GO	0.0039
MA	0.0798
MT	0.0582
MS	0.0785
MG	0.0019
PA	0.1736
PB	0.0094
PR	0.0015
PE	0.0503
PI	0.0569
RJ	0.0340
RN	0.0083
RS	0.0248
RO	0.1336
RR	0.4002
SC	0.0029
SP	0.0220
SE	0.0115
TO	0.4883
DF	0.1121

Table 1: Relative Errors with $\epsilon = 0.01$

State	Relative Error
AC	0.0081
AL	0.0005
AP	0.0229
AM	0.0009
BA	0.0000
CE	0.0007
ES	0.0011
GO	0.0003
MA	0.0027
MT	0.0010
MS	0.0002
MG	0.0001
PA	0.0015
PB	0.0000
PR	0.0001
PE	0.0003
PI	0.0013
RJ	0.0004
RN	0.0007
RS	0.0003
RO	0.0003
RR	0.0042
SC	0.0001
SP	0.0001
SE	0.0003
TO	0.0012
DF	0.0006

Table 2: Relative Errors with $\epsilon = 1$

References

[1] C. Dwork. Differential privacy. *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, 4052(2):1–12, 2006.

[2] C. Dwork. Differential privacy: A survey of results. *International Conference on Theory and Applications of Models of Computation*, pages 1–19, 2008.