Zayvion Worley

Dr. Viviane Nguyen

MTH 105

12/12/2024

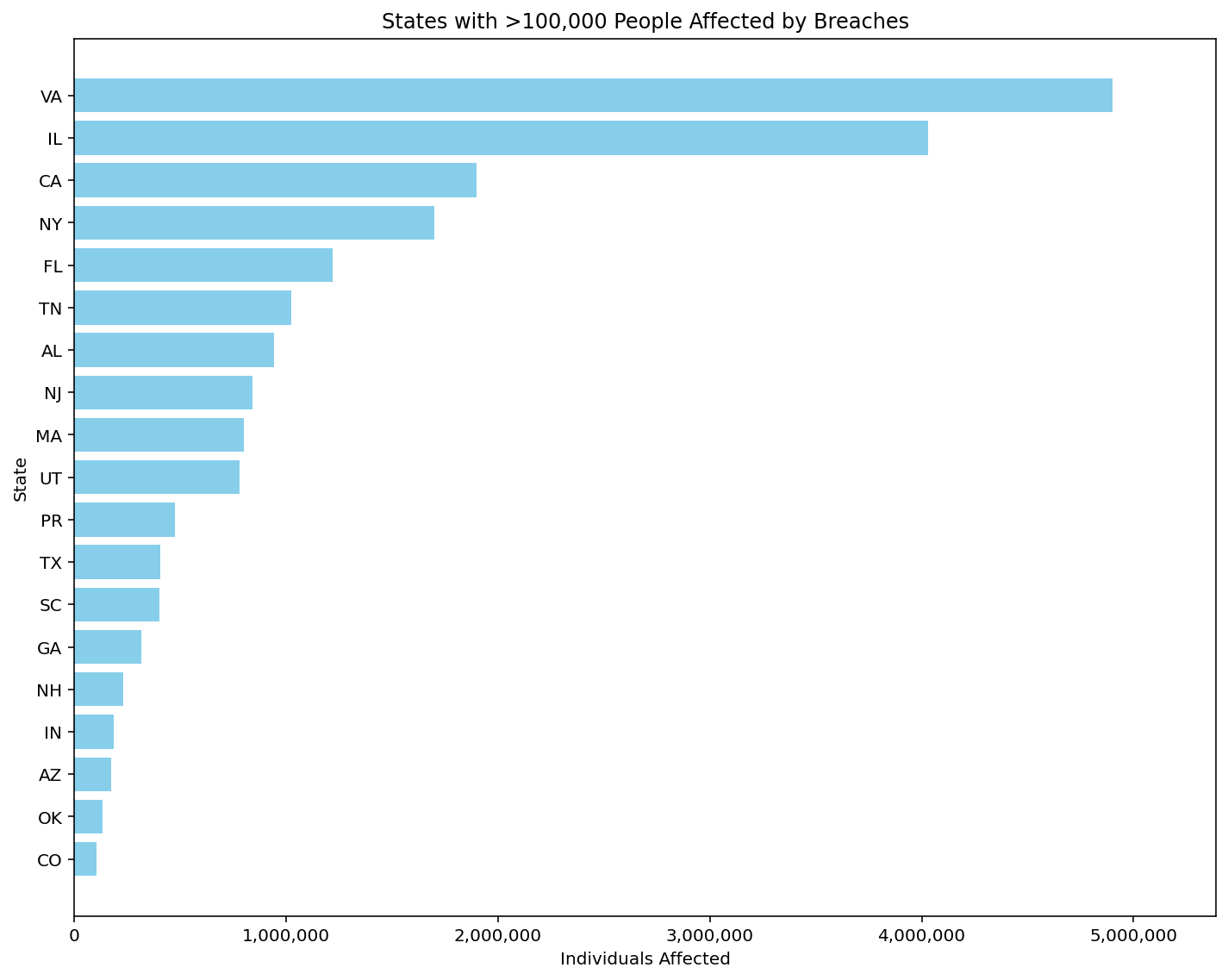
Cybersecurity Breaches: State, Breach Type Correlation and Other Analyses

According to the National Institute of Science and Technology, a breach is defined as “the loss of control, compromise, unauthorized disclosure, unauthorized acquisition, or any similar occurrence.” In this paper, I mainly want to figure out whether or not there is a correlation between the Breach Type and the State that the breach occurred in, but I will also be answering other questions as well. The data set that I will be performing analysis on tracks these breaches across different states, times, breach methods and locations.

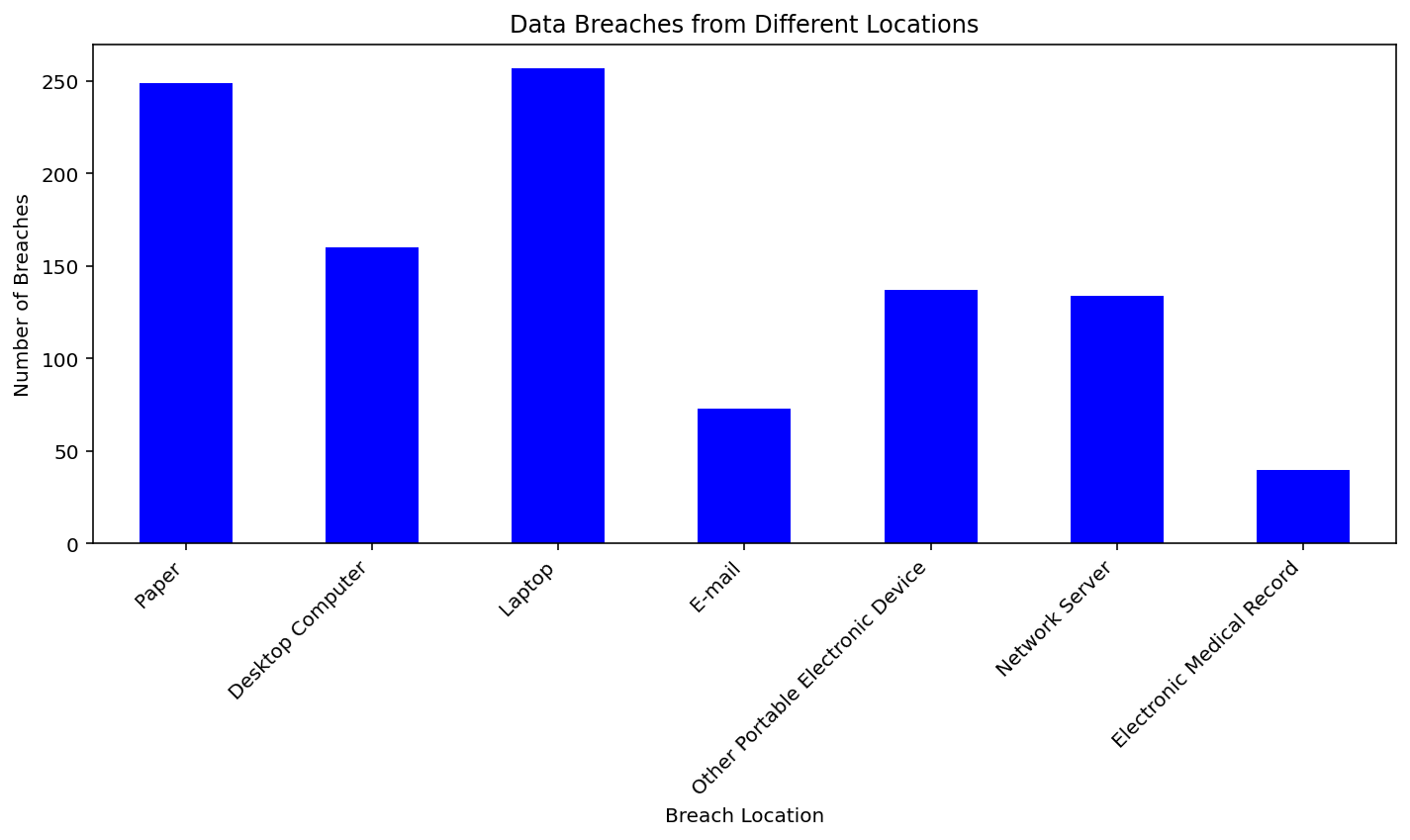
Let’s take a closer look at the dataset itself. It has 1055 total entries or rows, and 14 different columns. These columns are: 'Name\_of\_Covered\_Entity', 'State', 'Business\_Associate\_Involved', 'Individuals\_Affected', 'Type\_of\_Breach', 'Location\_of\_Breached\_Information', 'Date\_Posted\_or\_Updated', 'Summary','breach\_start', and 'year' As for the ‘scope’ of the data, it only records data breaches that occurred from 10/16/2009 to 5/27/2014, so there is a potential for the data to be outdated.

One interesting takeaway that I gleaned from the data was the mode type of breach, or the most common type of breach across the entire dataset, which happened to be theft. This makes sense, as most cybersecurity threat actors are interested in data theft. Another interesting takeaway is that the amount of breaches that involved a business associate was 25.69%. This implies that most data breaches do not involve a third party, which means that the reason these organizations were breached in the first place is not as simple as them working with an unscrupulous business associate. Interested in getting a better picture, I checked to see if the average amount of individuals affected across different types of breaches varied enough to tell me what types of breaches affected the most people. Interestingly enough, the ‘Unknown’ type had the most amount of people affected, meaning that the most devastating attacks don’t even have a clear cause, implying some need for increased surveillance and documentation so that future cybersecurity researchers can get a better understanding of how breaches occurred.

Now, it is time to discuss the columns of interest and the visualizations I drew from them. One column of particular interest to me is the ‘Individuals Affected’ column. This column is interesting because it shows just how many people are affected by each data breach, giving a sense of gravity to the analysis of the data as many people’s livelihoods could be affected by these data breaches. One thing I noticed immediately was the fact that the values varied wildly; while the mean was only 30262, the standard deviation was 227860. When the standard deviation is larger than the mean, this suggests that there is great variance in the data. This revelation is what made me avoid trying to check the correlations involving the number of individuals affected, as the values varied too wildly to provide a neat plot. Using this data and another column, ‘State’, I was able to create a histogram that orders the states by the amount of breaches that happened within them.

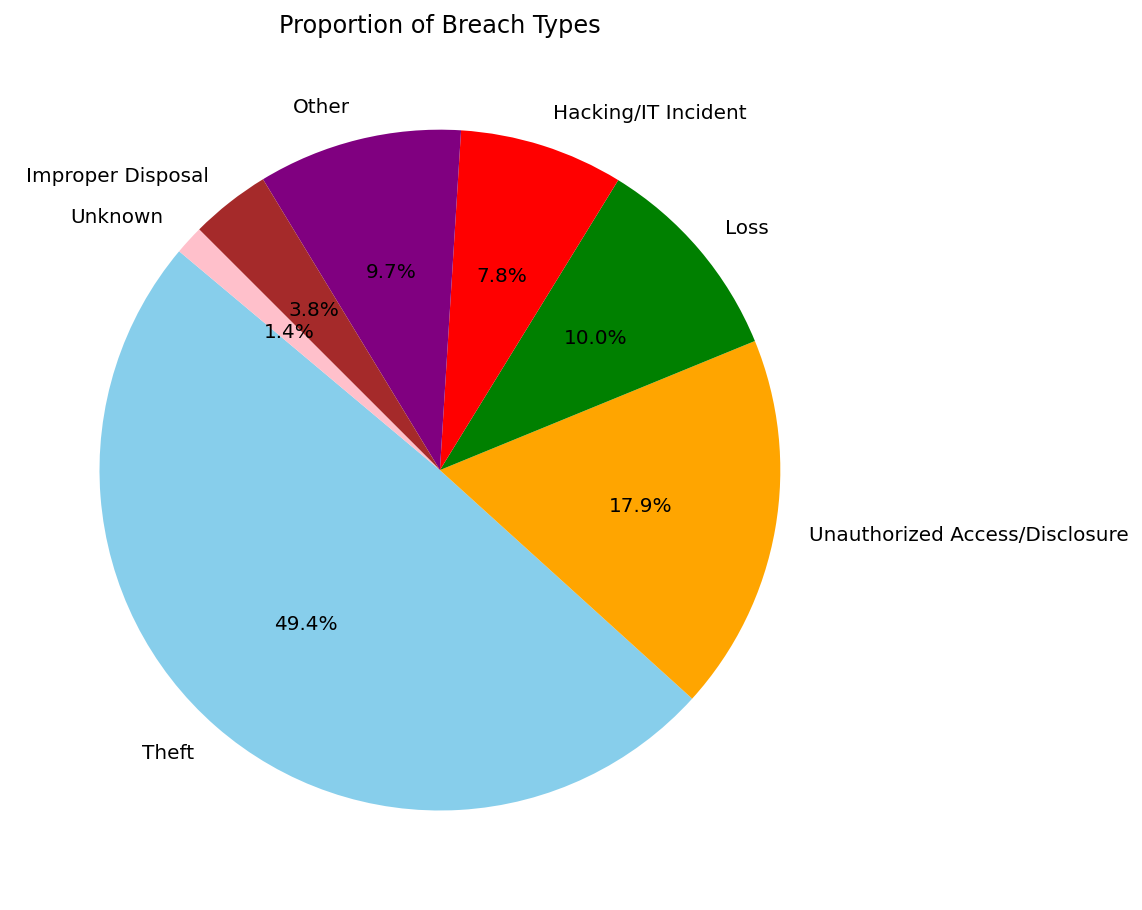


Due to the sheer amount of states recorded in the data I had to limit it to only states where >100,000 people were affected in breaches. The data clearly shows that the amount of people affected varies wildly between states, with Virginia, Illinois, and California standing at the top of the list. Virginia has Washington D.C. which makes sense, but the other two are mysteriously high in the amount of individuals affected by data breaches. Perhaps it has something to do with their population.

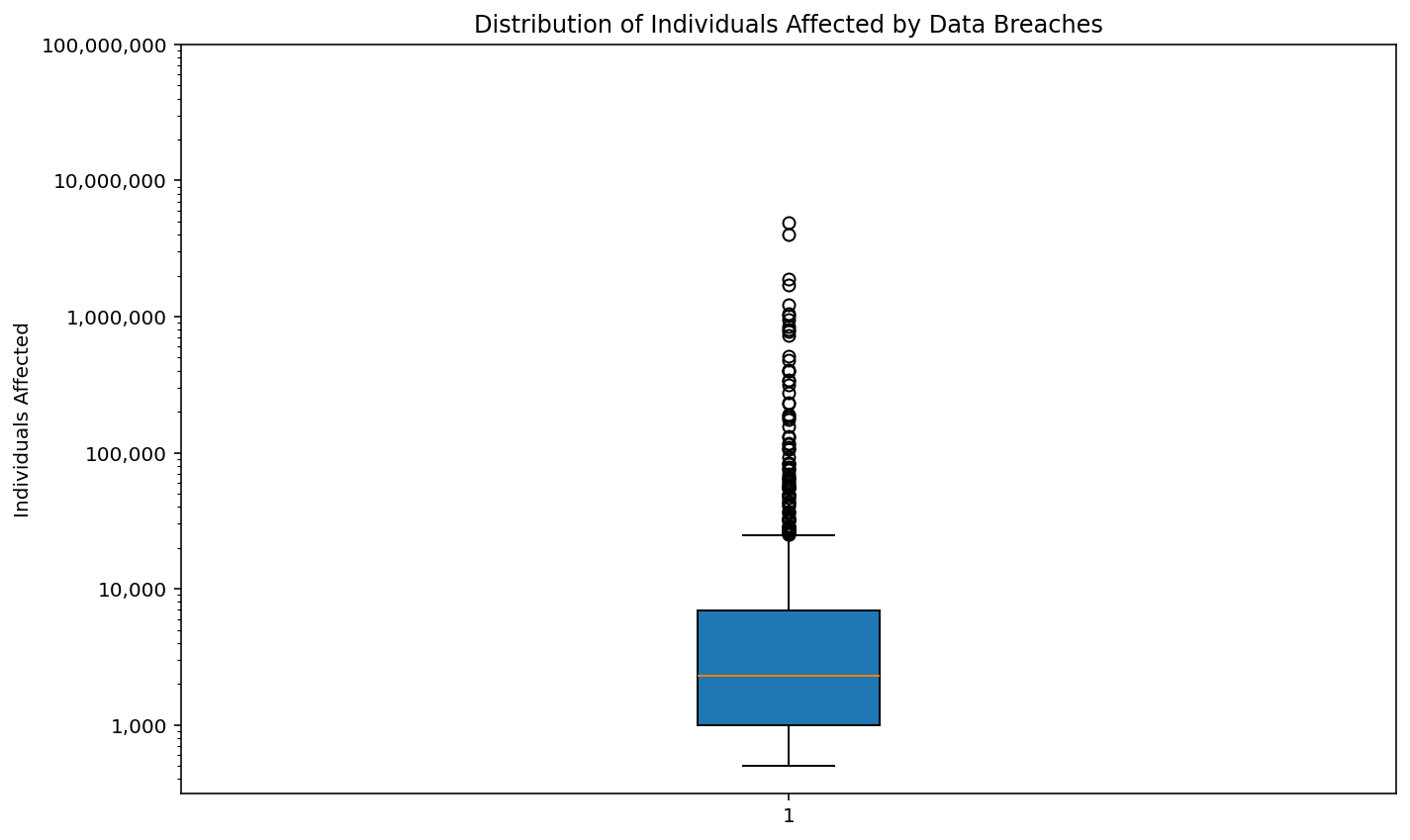
Another column of interest to me was the ‘Location of Breached Information’ column. It gives us important insight as to where an organization might want to shore up their defenses, change their policies and practices or begin investigating for potential threat actor activity. I went through the locations to see which locations were the most popular, and I saw that it was Laptop, with 257 total breaches. This means that laptops are at the most risk of being a weak point within an organization’s infrastructure. I created a bar plot using python to look at all the other breach locations and see how many times they occurred compared to the Laptop breach location.

The bar plot clearly shows that Paper is the second most common breach location, trailing just behind Laptop. This shows that another major reason organizations get breached is because of poor disposal of paper; a ‘paper’ breach shows that information was leaked through a method known as dumpster diving where a threat actor searches through trash to find inappropriately disposed of paper that may contain things like passwords and usernames. It could also simply refer to people placing notes on or around their workspaces with such information on them, making it very easy for an attacker to extract information. The least common type of breach was Electronic Medical Record. This is likely because medical records have specific laws around them that encourage organizations to be very vigilant in their defense, lest they pay the expensive legal fees that come with having that data breached.

The third column that I decided was interesting to me was the Type of Breach. This column got into the details of how the breach actually occurred. Some of the breaches had multiple different types involved separated by “, “, so I had to account for that in the code I used for the analysis. A few of the possible entries recorded by the data are Theft, which details the stealing of data for potential sale or criminal use, and Unauthorized Access/Disclosure, which details someone getting into a place that they shouldn’t or simply being told something that they shouldn’t and using that access or disclosure to breach a system and steal information, a more ‘physical’ type of data breach that could easily be solved but also easily overlooked. I created a pie chart so that I could compare the percentages of different occurrences and I found that Theft and Unauthorized Access/Disclosure were the top two most popular ones, suggesting a need to physically secure data and increase employee training such that they don’t simply give out important information that an attacker could use.

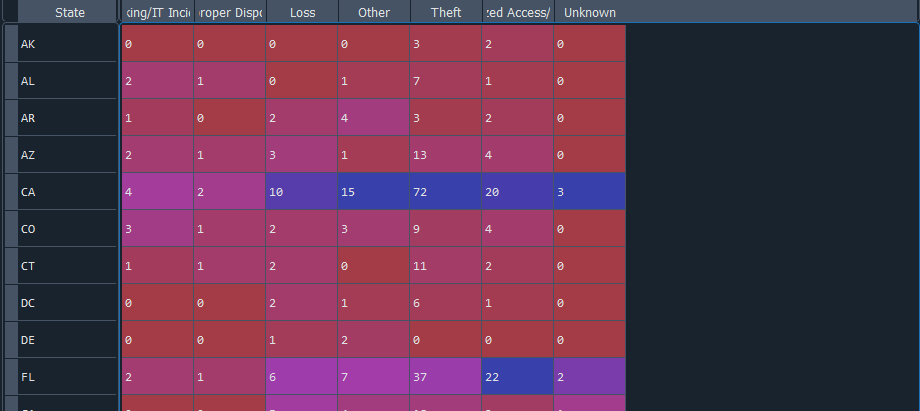


The least common types of breach are clearly improper disposal and unknown at 3.8% and 1.4% respectively, meaning that focusing on disposal methods is not necessarily a priority, and that most of the time we know the particular method through which data was breached which is a good sign, reflecting the efforts of IT staff to properly understand the breaches that occur in their organizations.

Going back to the subject of individuals affected, I wanted to see just how much variance that column had among all of the entries in the dataset, so I constructed a boxplot. It clearly reflects the five-number-summary I did earlier in the dataset, especially when I compared the mean and the standard deviation.

In the box plot, there are many outliers that lie outside of the range of the whiskers, suggesting extreme variance and implying that it may be inappropriate to run any other tests on it. The numbers varied so much that I had to make the Y axis logarithmically scaled so that it wasn’t incomprehensible, but still, it shows us that the amount of people affected per breach can vary wildly depending on a number of factors that require further study and research.

Dissuaded by the variance in the Individuals Affected column, I instead looked to the Type of Breach column and decided to figure out whether the Type of Breach could have a correlation with the State that the breach occurs in. This requires a contingency table to be run, which involves two qualitative variables. A contingency table basically works by counting every time that two variables ‘intersect’, in this case we’re seeing how many times the state of a breach intersects with the type of breach to get an idea of the breaches that occur in individual states.



Interestingly the data from the contingency table reflects some of the conclusions we drew from the histogram; you can clearly see that California has many data breaches of all kinds which is likely the reason it is the state with the highest number of individuals affected. But it also differs as well, Florida wasn’t that high in individuals affected but it has many entries in the contingency table, which implies that the breaches in Florida are numerous but do not involve many individuals.

Now that we have the contingency table, we can perform a chi-square test. A chi-square test is a statistical test used to determine whether there is a significant association between categorical variables or whether observed data significantly differ from expected data. It is often used in hypothesis testing to evaluate goodness of fit, independence, or homogeneity. The chi-square test is essentially focusing on seeing whether you can prove or disprove the null hypothesis. The null hypothesis says that there is no relation between the variables that are being compared, so to disprove the null hypothesis is to say that there is a relation. There are four values that a chi-square test returns: Firstly, it returns the chi square statistic. The chi square statistic is a numerical measure used to compare observed data with expected data under the null hypothesis. It quantifies the difference between observed and expected frequencies in the data. A small value indicates that the observed data closely matches the predicted data, while a large value indicates that there is a significant difference between the two which weakens the null hypothesis. Here, it is clearly 400.31, which is considered large, so there is evidence that the null hypothesis can be denied. Secondly, it returns the Degrees of Freedom which is important for determining the P-Value. Finally, it returns the P-Value, which is the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true. If the P-Value is lower than the significance level (in this case, 0.05), then the null hypothesis can be disproved and thus the values are related. If it is higher, then the null hypothesis cannot be disproved and the values are not related. This chi-square test returned a P-Value of 0.00023, which is lower than 0.05. Therefore, there is a correlation between the state that you find yourself in and the type of breach that those within the state will experience.

This paper’s findings are significant. They tell us that certain states have massive numbers of people affected by data breaches, that the amount of people affected by data breaches can vary wildly, which types and locations of breach are most and least common, and they confirm a link between the state a breach occurs and the type of breach that those in the state will experience. Further research will be needed to determine why these statistics show up the way that they are, but it is clear that certain states will need to shore up their defenses, and organizations everywhere will need to develop specific countermeasures against the most common breach types and locations outlined here.

References

The Devastator. (2022). Data Breaches: 30,000 Records of cyber-security data breaches (Version 2) [Data set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/data-breaches-a-comprehensive-list/