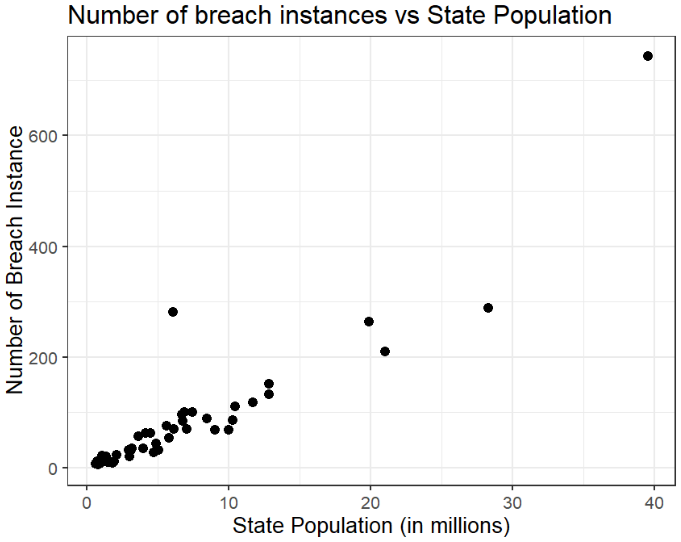
Executive Summary

Michelle Chen, Yimin Wang, Jing Yi Zhou

Since 2005, there has been an explosion of data being produced and aggregated as a result of technological advancement. It makes sense that as the amount of data grows, so does the number of opportunities for malicious entities to expose and take advantage of our personal data. This beckons the discussion of how companies and organizations can better protect themselves from data breaches. In our exploratory and data analysis project, we sought to explore various regional, industry-based, and temporal patterns that will help to shed light on data breaches in the US and what types of entities are at risk.

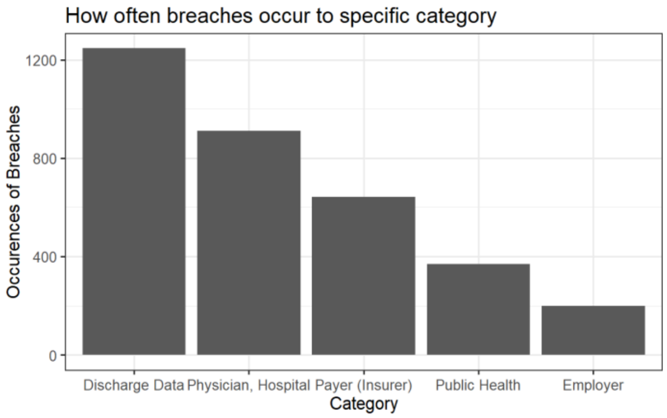
Looking at the relationship between 2017 population size and number of breaches, we see a clear positive linear relationship, which is what we would expect to find: as population size increases, the number of breaches increase as well. Other initial findings using time series visualizations also confirmed that the number of breach instances per year increases over time. This aligns with our expectations that as the amount of data being produced increases, entities are at greater risk for a potential compromise with their data.

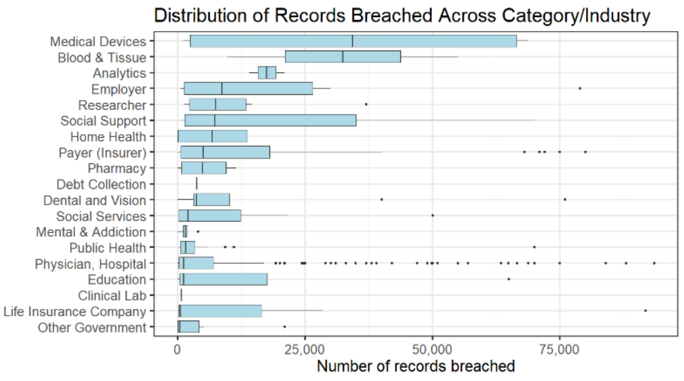


We did, however, discover two potential outliers: Maryland and California both experienced more breaches than expected. A potential reason for these outsized number of breaches for California and Maryland could be because California is a big tech hub with a huge population and Maryland is a big hub for health care technologies, which would lend for increased attempts at data breaching.

We also wanted to observe which industries were more likely to experience data breaches. We found that discharge data, payer(insurer), physician/hospital, public health, and employer were the top 5 categories/industries that were impacted by data breaches. When combined with our findings on the top 5 industries with the highest median records breached, we discovered that certain industries such as Public Health that are susceptible to data breaches may not necessarily have large magnitudes of data compromised per breach.

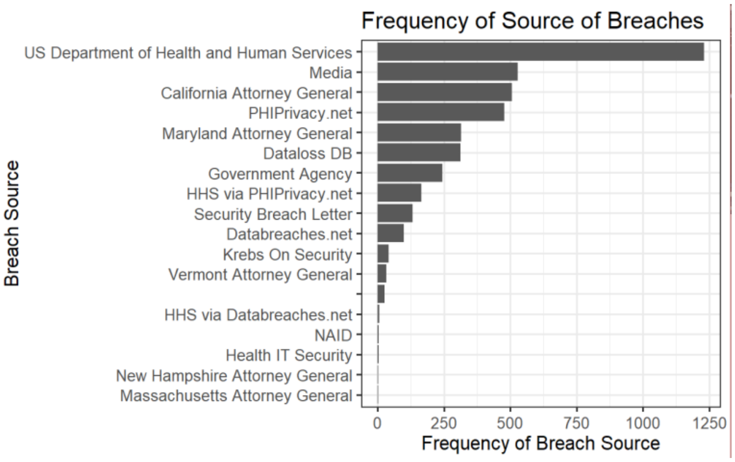
Meanwhile, we also noted that each industry's breaches can be attributed to a unique breakdown of various breach types depending on the context of the industry. For instance,

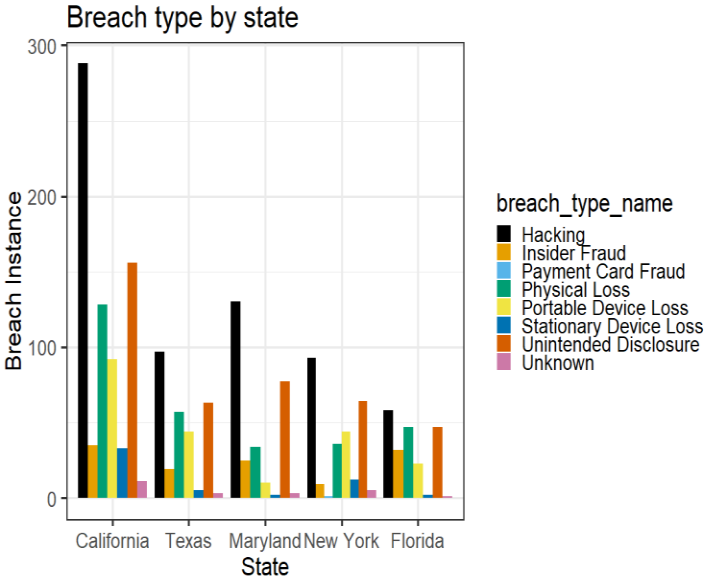




comparison of a health-related category Blood and Tissue with the Payer (Insurer) one, we saw that while Blood and Tissue experienced more insider fraud related incidences and no hacking breaches, the opposite held true for Payer (Insurer).

Deciding which outlets to publicize breach information through is another interesting facet of the data we looked at. Having a US government agency and Media in the top 5 most common breach sources was somewhat expected because the public does receive a lot of breaking news via media outlets (ie. news, social media, etc.) or press releases from the government. However, it was surprising to see the CA and MD states' attorney generals in the top 5 as well. Even more of an fascinating insight was that Texas, which was runner up in the data breach frequency across states visualization, and other states with high occurrences of breaches did not have their attorney general featured in our top 5 common breach sources. We speculate that there are certain areas that the attorney general of particular states will present on depending on how relevant it is to their specific function.





Upon further examination of the types of breaches that are the most prevalent within the top 5 states with the highest number of breaches, there seemed to be an overarching trend with hacking, unintended disclosure, and physical loss as the major reasons for breaches. Knowing this information, companies in certain states can take preventative measures to secure their data against these potential threats.