**FALL 2023**

**INFO 5709**

**DATA VISUALIZATION AND COMMUNICATION**

**Final Project Report**

**Analysis of Healthcare Data Breaches**

**Srinivas Pallapu**

**11604461**

**INTRODUCTION**

Healthcare is one of the most critical sectors with involves a lot of personal and sensitive information making it a primary target for cyber-attacks. Healthcare data breaches are believed to be a significant threat to the privacy and security of individuals' information. In the United States, the Department of Health and Human Services Office for Civil Rights documents such incidents. The dataset used in this project has data on incidents up to September 14, 2021, containing details of various data breaches in different states. Different incidents varying from hacking to unauthorized access are included in the dataset to highlight the vulnerabilities of different healthcare entities and related services. One of the most contributing factors of this dataset is the column containing the count of individuals that were affected by the particular incidents.

The project will explore and identify the most affected entities, breach incidents, and some other useful insights that could help take proper actions by providing appropriate resources in healthcare to reduce the cyber-attacks and prevent any loss of health-related data.

**DATASET DESCRIPTION**

The dataset used in this project is titled “Healthcare Data Breaches” and contains 866 records of different data breaches in the healthcare domain. The dataset contains 9 columns. Some data records and columns have missing values which need to be handled before doing the data visualization in Tableau.

One of the columns in the dataset “web\_description” is empty for the entire dataset and thus, it has been removed from further consideration. The “States” column contains 4 missing values, which are replaced by Unknown to avoid any misconception use of replacement with most occurring state.

Below is a brief description of each of the columns in the dataset:

1. **Name\_of\_covered\_entity**:
   1. Data Type: String
   2. The name of the covered entity, i.e., the organization or entity responsible for protecting the healthcare data.
2. **State:** 
   1. Data Type: String
   2. The U.S. state where the data breach occurred.
3. **Covered\_entity\_type**:
   1. Data Type: String
   2. The type of covered entity affected by the breach (e.g., healthcare provider, health plan, business associate).
4. **Individuals\_affected:** 
   1. Data Type: Integer
   2. The number of individuals whose data was affected or compromised in the breach.
5. **Breach\_submission\_date:** 
   1. Data Type: Date
   2. The date when the breach was reported or submitted.
6. **Type\_of\_breach**:
   1. Data Type: String
   2. The nature or type of the data breach (e.g., unauthorized access, hacking, loss of physical records).
7. **Location\_of\_breached\_information:** 
   1. Data Type: String
   2. Details about where the breached information was stored, or the specific system affected.
8. **Business\_associate\_present**:
   1. Data Type: String
   2. Indicates whether a business associate was involved or present in the breach. It could be a "yes" or "no" value.

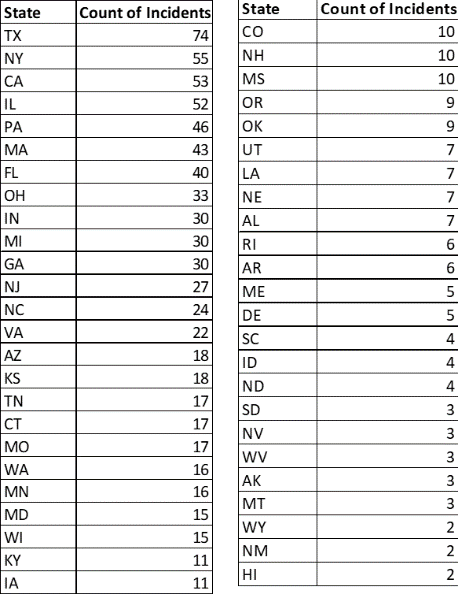
**TOOLS USED IN THIS PROJECT**

* Python Programming
* Microsoft Excel
* Tableau

**EXPLORATORY DATA ANALYSIS**

The exploratory data analysis of the dataset used for this project was done using Python programming and MS Excel pivot tables. Multiple columns were examined to get some of the numerical statistics and distinct values of some of the categorical variables.

1. **Number of Incidents in different states**.



A graph of colored lines

Description automatically generated with medium confidence

The distribution of healthcare data breaches across different states provides information about the geographical disparity in incident counts. Texas (TX) is leading in the list with 74 incidents, next in the list is New York with 55 incidents, California with 53 incidents, and Illinois with 52 incidents. These 4 states make up a majority of 27% of all the incidents across the United States.

Hawaii, New Mexico, Wyoming, and Montana are some of the states with fewer data breach incidents (less than 5).

This difference in the count of incidents can be due to different densities of healthcare providers, varying cybersecurity measures, population, and crime rates in the states.

1. **Count of different Types of covered entities affected by the data breaches.**

From the bar plot visualization, we can see the healthcare providers with a count of 565 are the most impacted organizations showing their vulnerability directly as they are involved directly in patient care or patient data.

A graph with numbers and a bar

Description automatically generated with medium confidence

1. **Types of Breach**: Based on the below visualization, it is evident that majority of the healthcare data breaches are because of Hacking or IT-related incidents. Unauthorized Access accounts for the next type of breach.

A graph with different colored squares

Description automatically generated

1. **Most Vulnerable Location of Information based on the attack’s history.**

As in the previous visualization (3), Hacking or IT-related incidents were highest among the data breaches in the healthcare industry and the below snippet shows the count of the location of information that was attacked during the breaches. Network Server and Emails are the two most vulnerable areas by which the attackers make attempt to launch their cyber-attacks.

A screenshot of a computer

Description automatically generated

**HYPOTHESIS**

1. Does any particular geographical location like States have a significant impact on the number of affected individuals by data breaches in the healthcare sector?
2. Are there any specific types of data breaches more prevalent in some of the states only?
3. Has there been a significant increase in the frequency of healthcare data breaches over the last 3 years i.e. from 2021 to 2023?
4. Do certain types of healthcare data breaches exhibit a correlation based on the location of the information storage that was breached?

**TABLEAU GENERATED VISUALIZATION**

1. The first hypothesis was to determine if the different states in the USA have a significant number of affected people due to the data breach incidents. The below two visualizations created in tableau, Geographical Map and Vertical bar plots show the count of individuals affected in each of the states.

A map of the united states

Description automatically generated

In Geographical Map above, the darker shade of blue highlights the higher number of affected individuals whereas the light shade is for states with less affected individuals due to the healthcare data breaches.

Below provided vertical bar plot, adds value to the previous geographic map by showing the states with count of affected individuals in a sorted order. Colorado and Tennessee are the states which have the most affected individuals due to healthcare breaches.

A graph of a number of individuals affected by data breaches

Description automatically generated

1. The data breaches in the dataset are identified as hacking/IT incidents which are most common as already seen in most of the states. Coming to other types of data breaches of Improper Disposal of Data, Loss of Data due to negligence and Unauthorized Access can be seen prevalent in particular states shown in the below visualizations.

Improper Disposal was found particularly in the state of Texas and 1 in New Jersey and Illinois.

Loss due to negligence was 3 incidents in Illinois, and 1 each in New York and New Hampshire

Breaches due to threat incidents are common in the state of California.

Hacking and IT-related issues are prevalent in all the states of USA in any healthcare data breach.

A screenshot of a graph

Description automatically generated.

A graph of a bar graph

Description automatically generated with medium confidenceA graph with blue squares

Description automatically generated

1. The fourth hypothesis was focused on identifying the increase in the number of incidents of particular data breach types over the last three years.

From the below line chart, there has been a significant increase in the number of hacking or IT-related incidents across the country. It increased from 11 incidents in 2021 to 267 incidents in 2022 and further almost doubled to 443 incidents in the year of 2023.

Unauthorized access-related incidents are newer in comparison to the other breaches. Maybe because of improved security measures that are now able to identify such elements of unauthorized access in different organizations.

The other types of breaches apart from hacking and unauthorized access have not seen an increase in the number of reported incidents, they have remain stagnant over the years.

A graph with a line and a line

Description automatically generated with medium confidence

Same with the number of individuals affected by the hacking increased significantly from 33,002,030 in the year 2022 to 104,406,689 in the year of 2023. This count nearly tripled within the gap of one year.

A graph of a number of individuals affected by data breaches over the years

Description automatically generated

1. The location of the information that was breached is critical in identifying the areas that required special attention or resources to be provided for safe keeping and preventing future attacks.

A blue and green squares with white text

Description automatically generated

The above tree map clearly shows the most common target of the attacks online are the network servers that store the critical health data records of the patients. This number was equal to 546 of the number of breaches out of 867 breaches in this dataset over the last 3 years. This accounts for about 65% of the total data breaches. The second target area is the Email in the hacking incident that comprises 137 data breaches.

**CONCLUSION**

This data analysis and data visualization project provides valuable insights into healthcare-related data breaches over the last three years of 2021 to 2023. We can conclude by stating there is a requirement for continuous surveillance and proactive measures to be taken by all organizations working in the domain of health. There has been a significant increase in the number of online hacking-related incidents reported this year combined in the entire USA and particularly in some states like California and Colorado.

Thus, an increase in the number of incidents leads to an increase in the number of affected individuals by such cases eventually risking the reputation of the organisation and losses in finances as well. There is a significant need to address cybersecurity challenges in the healthcare sector. With today's technological advancements, and with new cyber threat intelligence abilities organizations and government health agencies are making the stakeholders (either organizations or individuals) stay informed and avoid any such incident leading to loss of sensitive healthcare data. There is a need to investigate the loopholes and significant requirements of funding and investment to safeguard patient data and privacy.

**LEARNING FROM THE PROJECT**

From this project, I learned how data visualization can make simple data documents good enough to provide insights to business stakeholders and also to the common public who do not have any technical understanding of the domain. Data visualization can make creative solutions to be visually appealing.

Apart from this, I learned how to use Tableau and develop dashboards to show combined reports to all the sheets together and also make use of filters to make the dashboard interactive.