Project - Dataset between 2009-2019 for healthcare data breach
(https://data.world/zendoll27/biggest-healthcare-data-breaches-2009-2019)
(https://www.privacyaffairs.com/healthcare-data-breach-statistics/)

- the topic of your project
- why it's worth investigating (problem statement)
- your research questions
- methods for data analysis

Topic: US Healthcare data breach between year 2009 and 2019: Insight and Implications

Why the topic is worth investigating: Because of the sensitivity of the data and how healthcare professionals mismanage health data. Also, the clear attack on the health sector as seen from incessant attacks (increased by 2733% in those 10 years).

Research Questions
How many times did each organisation get hacked?
// The mean of organisation ?? hist
How does it affect individuals?
// The mean of organisation ?? hist

How many times do data types lose / hack a particular organisation?

Category

What was the central tendency of healthcare data breach by a year?
2 years 2009 - 2019

// 2009 to 2015,2019 mean or median
//

// standard
How many percent of the population is affected by healthcare data breaches?
??
When was the worst year for an overall number of healthcare data breaches?

// highest

Why is healthcare data so frequently targeted?

Medical devices connected to the network are often unsecured?

Methods for data analysis: Descriptive analysis

Feedback from Lecturer: Perform inferential statistics by taking the mean for some years maybe 2011 and 2015 and compare them using T-test.

Project Workflow

- Abstract
- Introduction
- Dataset
- Exploratory data analysis
- Data cleaning and transformation
- Hypothesis Testing
- Research Question 1 and 2
- Conclusions and discussion

Completed

Abstract

Data breaches in the US healthcare industry have been a persistent problem over the past decade. Between 2009 and 2019, numerous healthcare organisations experienced data breaches, resulting in the unauthorised access or disclosure of sensitive patient information. These breaches have had serious consequences for both patients and healthcare organisations, including financial losses, damage to reputation, and loss of trust.

T-test data analysis as a statistical technique was used to analyse the dataset, US healthcare data breaches between 2009 and 2019. It involved comparing the mean values of different variables in two groups of data to determine whether there is a significant difference between the groups.

This project used the t-test data analysis to provide valuable insights into the characteristics and consequences of healthcare data breaches, relationship between variables, as well as correlation between features.

Introduction

In this report, we are going to explore the dataset that has been titled ‘Biggest healthcare data breaches 2009-2019’ from the Privacy Affairs website which we have retitled as `data` for this analysis. Additional documentation can be found here:

< https://data.world/zendoll27/biggest-healthcare-data-breaches-2009-2019>

The dataset includes 7 variables and 2,641 observations:

- ‘Name.of.Covered.Entity’

- ‘State’

- `Covered.Entity.Type`

- `Individuals.Affected`

- `Breach.Submission.Date'

- `Type.of.Breach'

- 'Location.of.Breached.Information'

First, let's activate the 'dplyr' package to explore the dataset after cleaning it in excel and importing it into R.

RQ

Correlation between entity type/state and individuals affected

Which state has the highest occurrence and which has the lowest.