CIS 5270-01: BUSINESS INTELLIGENCE PROJECT

ABSTRACT

Breaches are widely observed in the healthcare sector and can be caused by many different types of incidents, including credential-stealing malware, an insider who either purposefully or accidentally discloses patient data, or lost laptops or other devices. The goal of this study is to analyse and develop comprehensive visualizations of the types of breaches that occur in the healthcare industry with respect to various factors.

NAME: SIDDHI UDANI

GUIDED BY: DR SHILPA BALAN

CIN: 307401801

TABLE OF CONTENTS

A. DATASET URL: The datasets for this topic has been taken from two separate sources, the links to both the datasets have been mentioned below:

1. https://www.kaggle.com/archangell/hipaa-breaches-from-20092017/downloads/hipaa-breaches-from-20092017.zip/1

This dataset has been downloaded from www.kaggle.com and has been originally sourced from **Department of Health and Human Services** containing 9 interesting columns and about 1701 rows for different variations of analysis and visualizations, which can further be used for improving security against such data breaches**.**

The Columns and their descriptions are as follows:

Size of the dataset: 380KB.

COLUMN NAME	DESCRIPTION	TYPE
Name of covered entity	Name of the entity/company who suffered the breach.	Plain Text
State	Name of the state where the breach occurred.	Plain Text
Covered entity type	Covered entities can be institutions, organizations, or persons who electronically transmit any health information in connection with transactions for which HHS has adopted standards.	Plain Text

Individuals affected	Count of the individuals affected due to the breach.	Number
Breach submission date	Date when the breach took place.	Date
Type of Breach	Classified breaches as per their types	Plain Text
Location of Breached information	Technical Location where the data breach took place.	Plain Text
Business Associate present	Determines presence of a person or entity conducting certain functions on behalf of a covered entity.	Plain Text
Web Description	Description of the web url that was breached.	Plain Text

2. https://breachlevelindex.com/data-breach-database

This dataset has been archived from a centralized, global database of data breaches named: Breach Level Index, displaying 8 columns and 1484 rows. It shows that data breaches are very much a growing threat for organizations. The number of records compromised is remarkable, considering the lengths many organizations go to in order to protect their data.

Size of the dataset: 709KB.

COLUMN NAME	DESCRIPTION	TYPE
Rank	Determines the Rank of the breach.	Number
Risk score	Determines the Risk score/Risk level of the breach.	Number

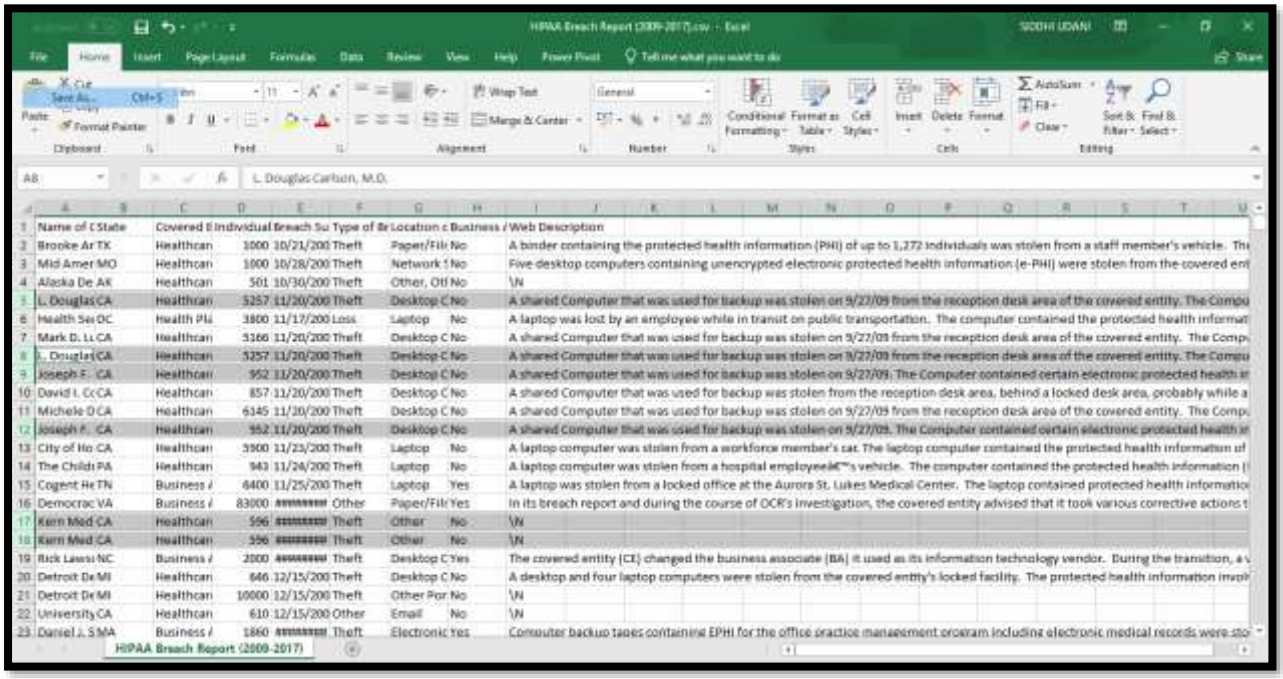
Industry	Shows the most common industries that suffer the breach.	Plain Text
Records Breached	States the number of records breached.	Number
Date of Breach	Shows the date when the breach occurred.	Date
Type of Breach	Tells us about the type of data breach that took place.	Plain Text
Source of Breach	Determines the source/reason of the actual data breach.	Plain Text
Location	States the location(country wise) where the breach happened.	Plain Text

B. DATA CLEANING:

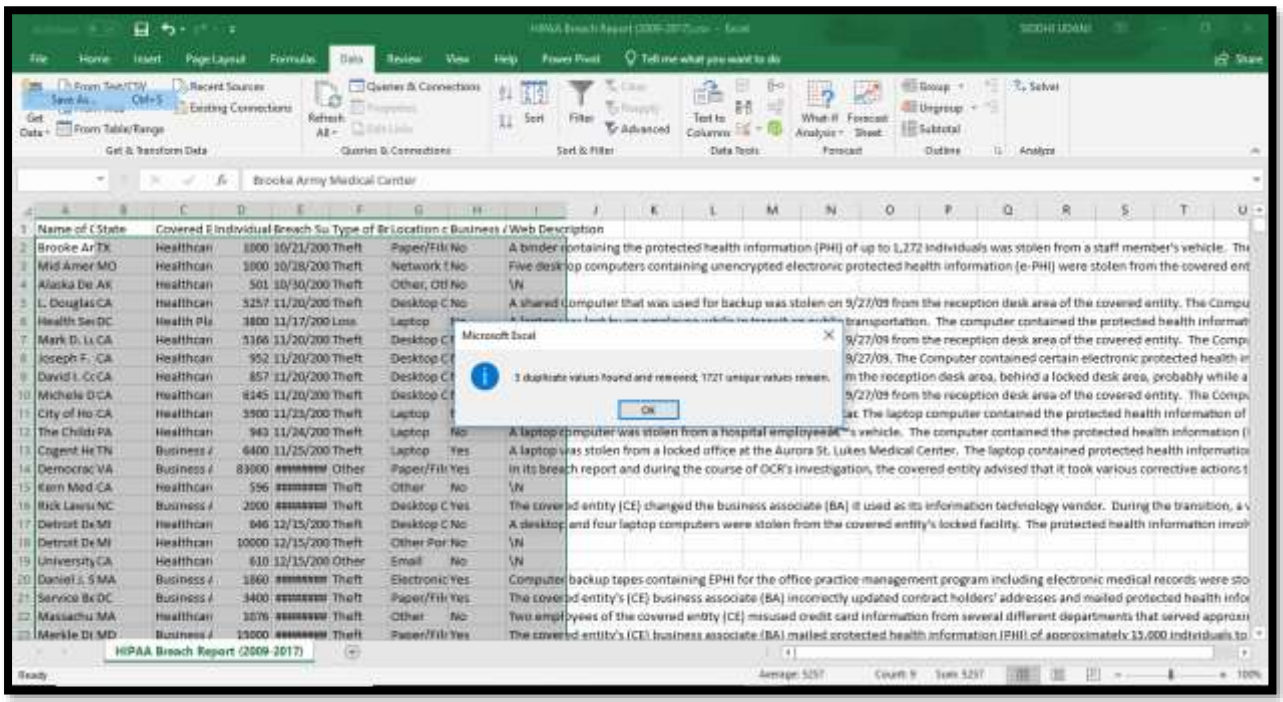
Data cleaning is especially required when integrating heterogeneous data sources and should be addressed together with schema-related data transformations. Data Cleansing or data scrubbing is the process of identifying and correcting inaccurate data from a data set. With reference to customer data, data cleansing is the process of maintaining consistent and accurate (clean) customer database through identification & removal of inaccurate (dirty) data. Here, inaccurate data stands for any data that is Incorrect, incomplete, out-of-date, or wrongly formatted.

1. **REMOVING DUPLICATE VALUES**: Since the dataset had a few duplicate values (duplicate rows), they were removed using "REMOVE DUPLICATE" option in "DATA" tab in Excel. The dataset had 3 redundant values, which were Removed and 1721 unique records were found.

➔ Before cleaning:

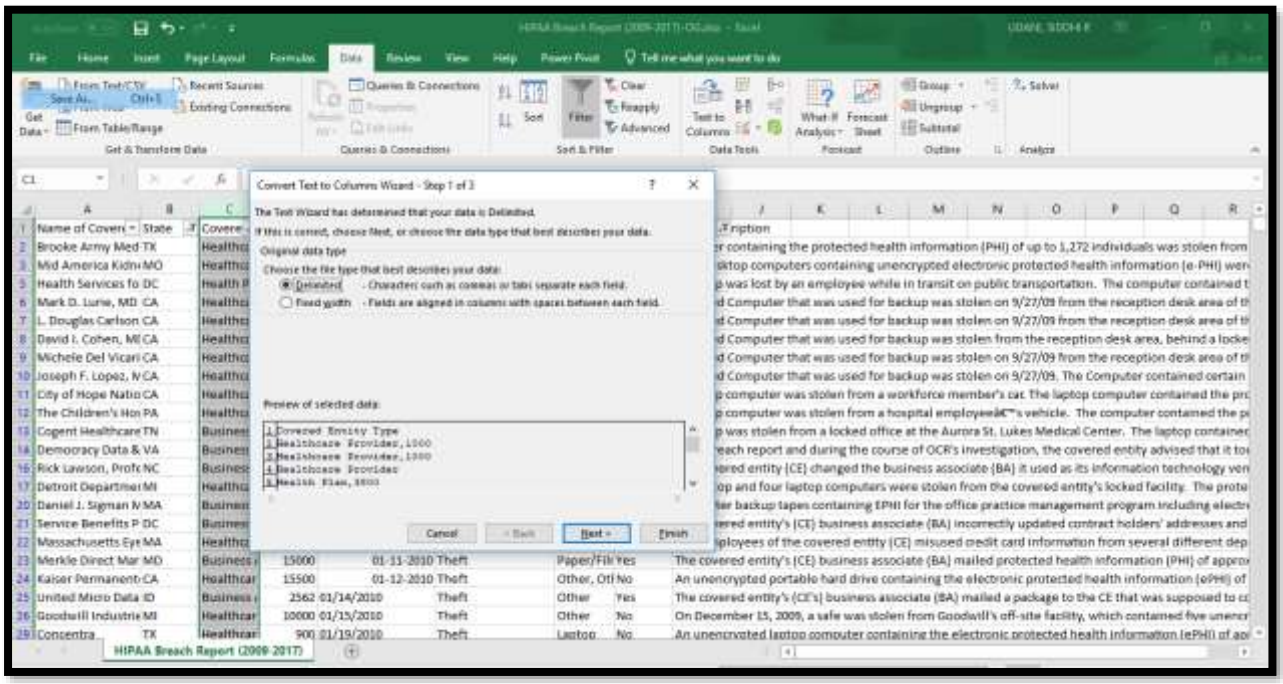


➔ After cleaning:

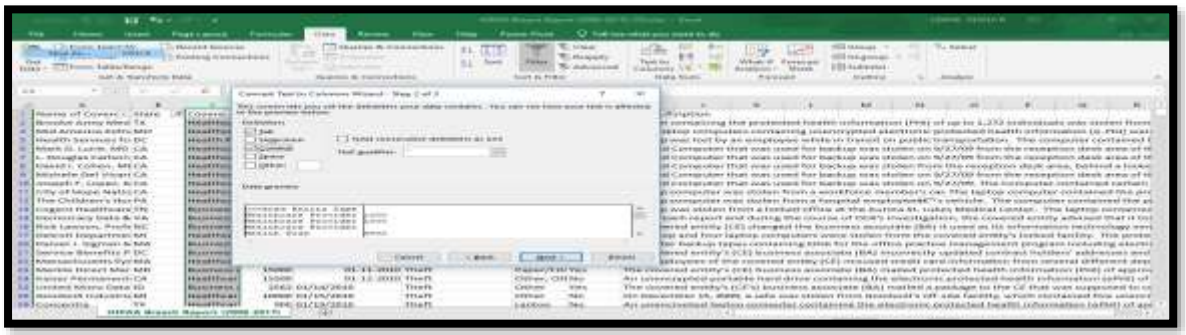


2. SPLITTING COLUMNS: Before there were two data in one column as shown in the picture. I separated these columns using programming feature “TEXT TO COLUMN” where characters such as ‘Tab’ and ‘Comma’ are used to seperate the fields in excel. So after separation as shown in the image, the ‘Covered entity type’ has been split into 2 following columns: ‘Covered entity type’ and ‘Individuals Affected’.

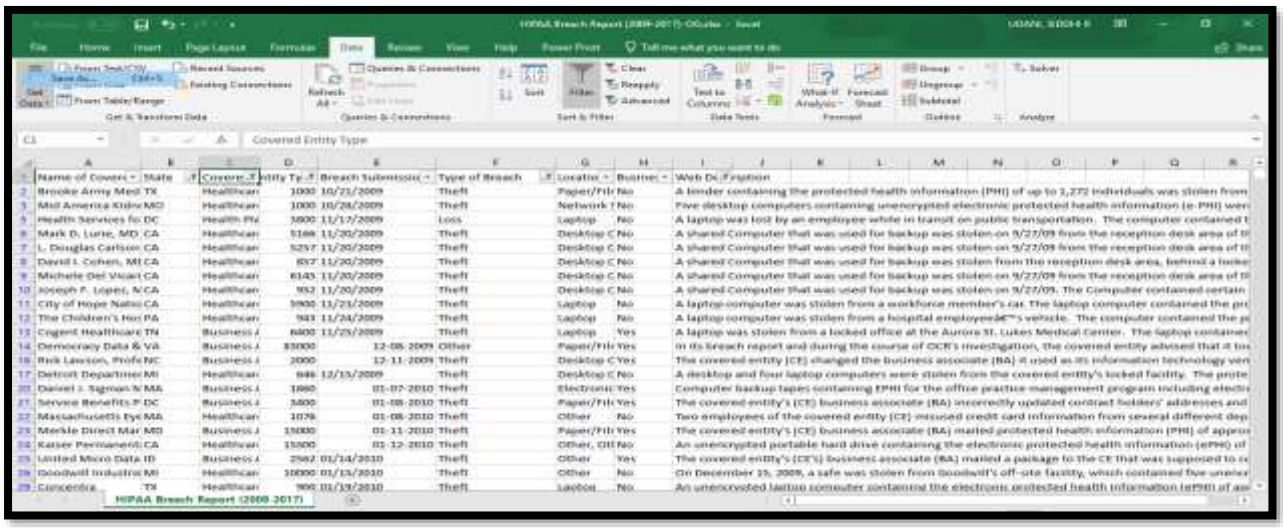
➔ Before cleaning:



➔ Cleaning Step:

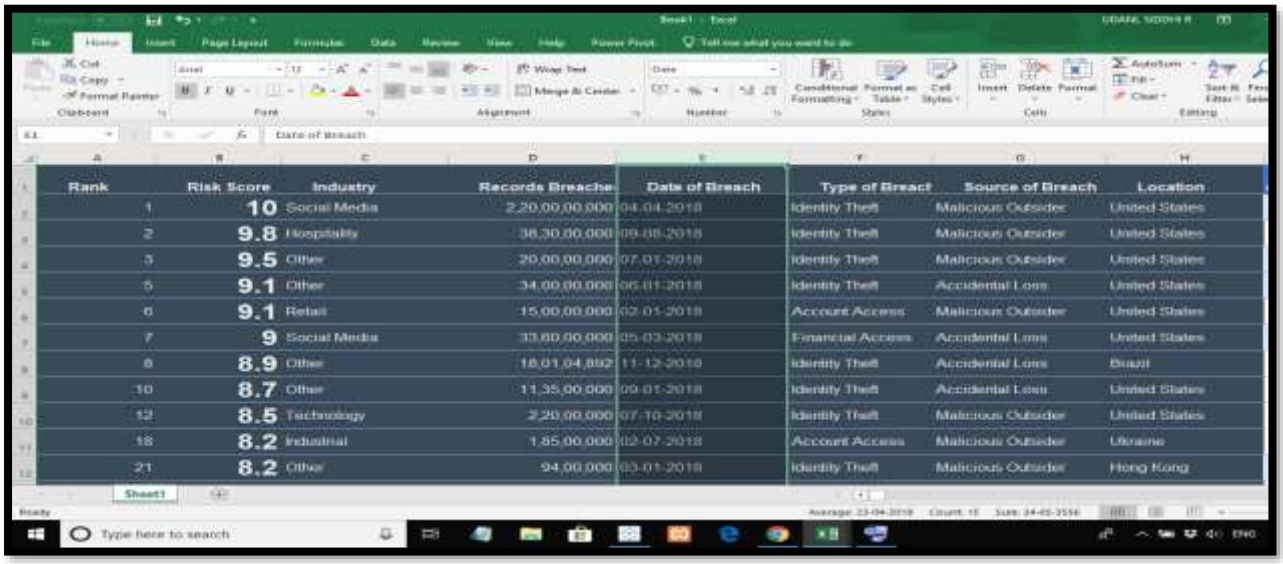


➔ After cleaning:



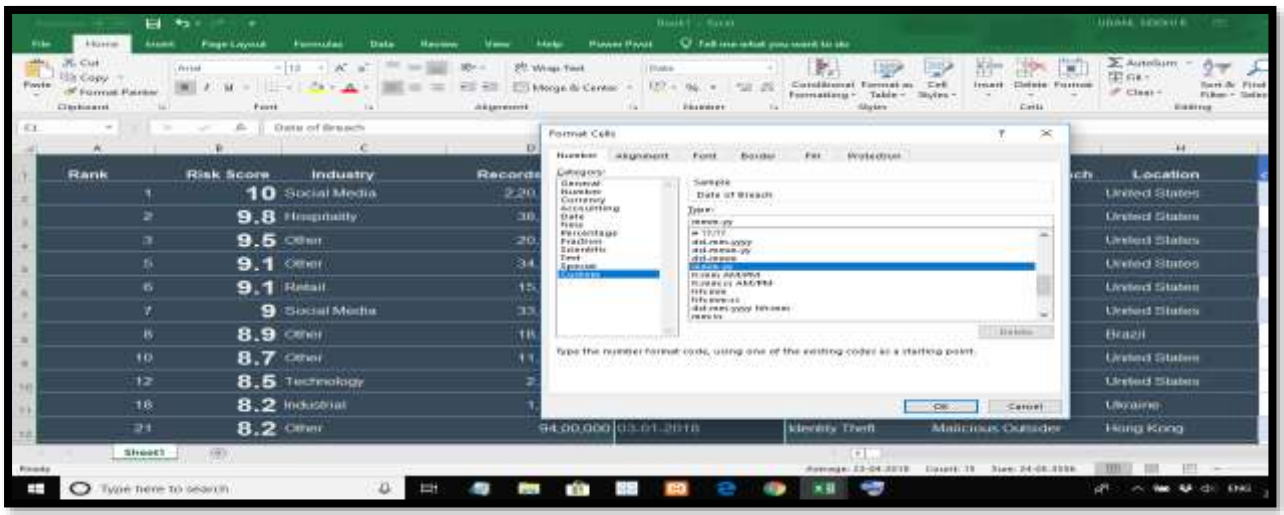
3. FORMATTING DATE FORMAT: Since the date format in my dataset was not properly cited as per needed, the dataset was formatted using the 'FORMAT CELLS' option in Excel, and after that was converted in the 'mmmm-yy' (month-year) format.

➔ Before cleaning:

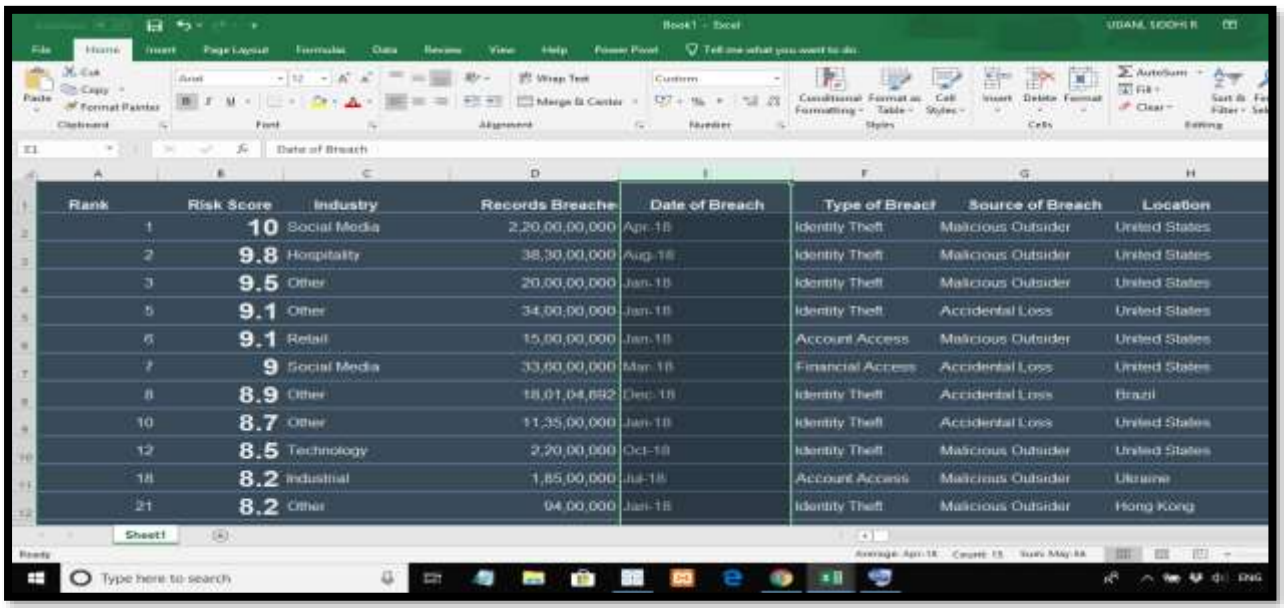


Rank	Risk Score	Industry	Records Breache	Date of Breach	Type of Breach	Source of Breach	Location
1	10	Social Media	2,20,00,00,000	04-04-2018	Identity Theft	Malicious Outsider	United States
2	9.8	Hospitality	38,30,00,000	09-08-2018	Identity Theft	Malicious Outsider	United States
3	9.5	Other	20,00,00,000	07-01-2018	Identity Theft	Malicious Outsider	United States
5	9.1	Other	34,00,00,000	05-01-2018	Identity Theft	Accidental Loss	United States
6	9.1	Retail	15,00,00,000	02-01-2018	Account Access	Malicious Outsider	United States
7	9	Social Media	33,00,00,000	05-03-2018	Financial Access	Accidental Loss	United States
8	8.9	Other	18,01,04,892	11-12-2018	Identity Theft	Accidental Loss	Brazil
10	8.7	Other	11,35,00,000	09-01-2018	Identity Theft	Accidental Loss	United States
12	8.5	Technology	2,20,00,000	07-10-2018	Identity Theft	Malicious Outsider	United States
18	8.2	Industrial	1,85,00,000	02-07-2018	Account Access	Malicious Outsider	Ukraine
21	8.2	Other	94,00,000	03-01-2018	Identity Theft	Malicious Outsider	Hong Kong

➔ Cleaning Step:



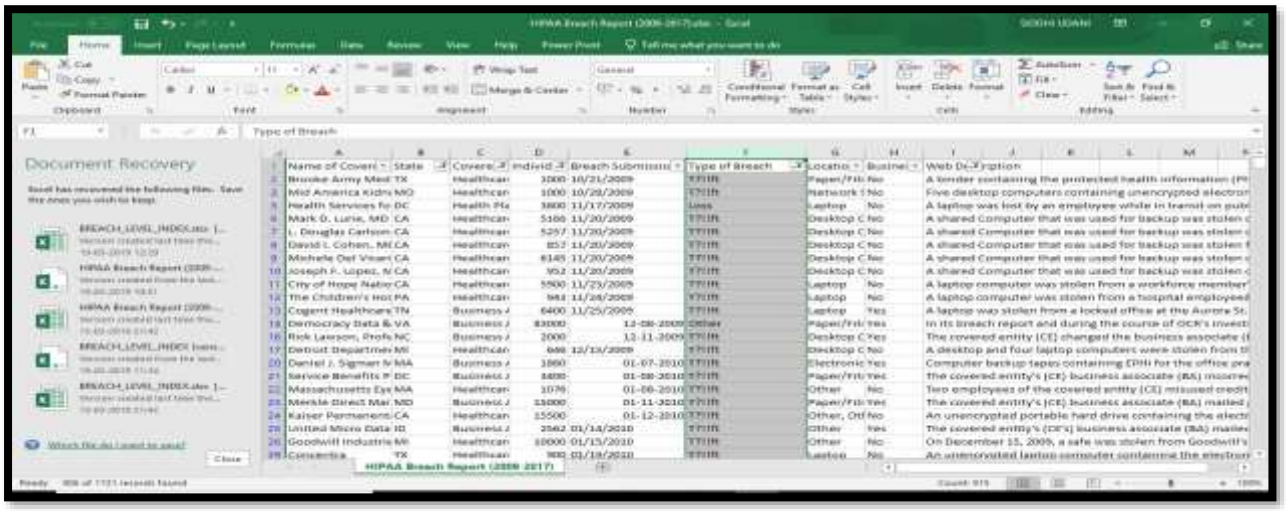
➔ After cleaning:



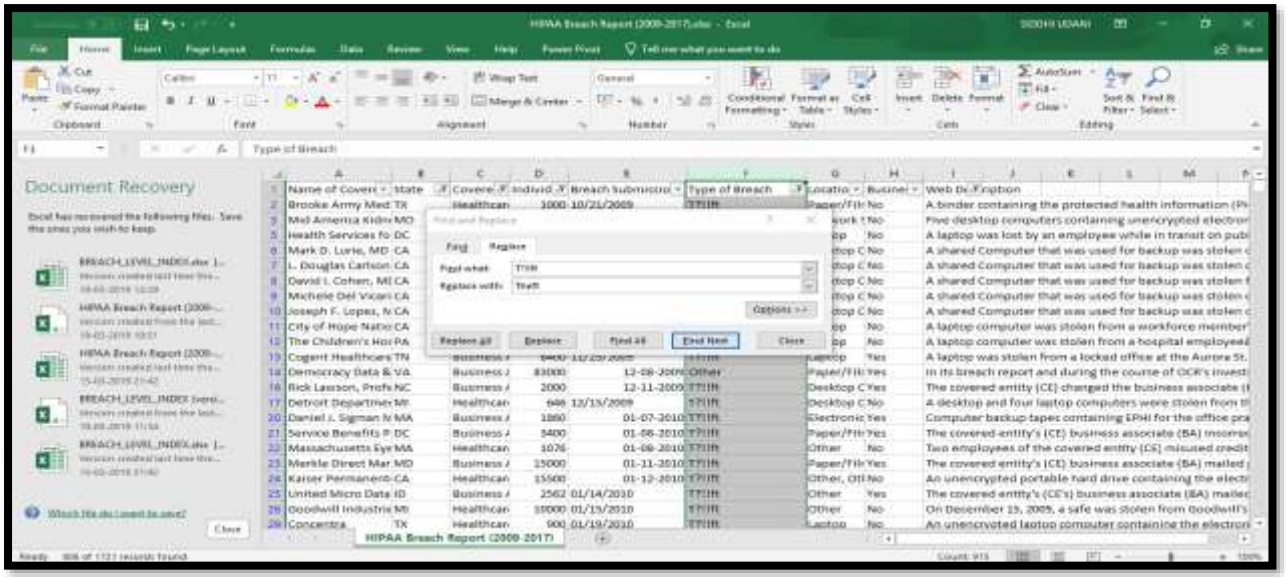
4. REMOVING CRYPTIC VALUES: The dataset contained a few cryptic values with the error which exists if any of the values has the wrong data type. So i removed

these values and replaced them with the actual word: “Theft” and with this method, 460 of such values were cleaned.

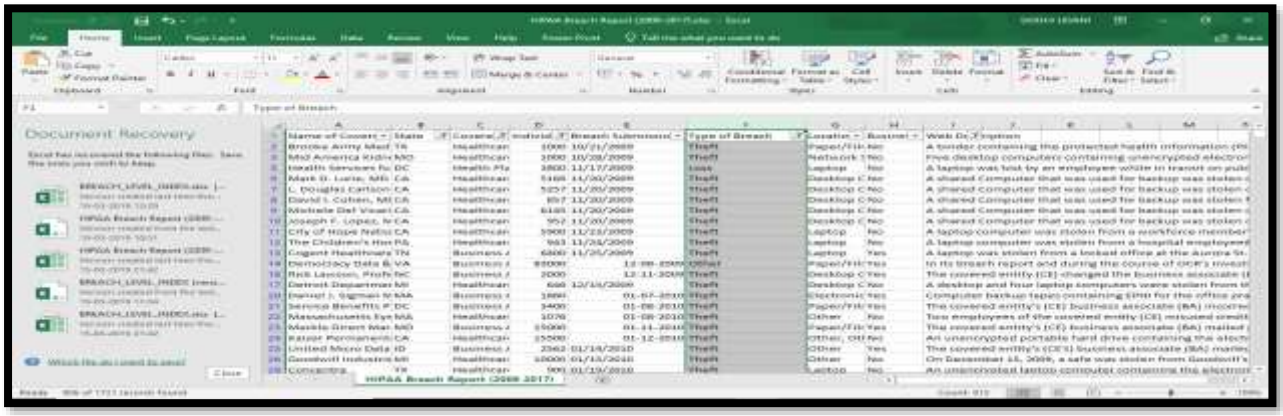
➔ Before cleaning:



➔ Cleaning step:

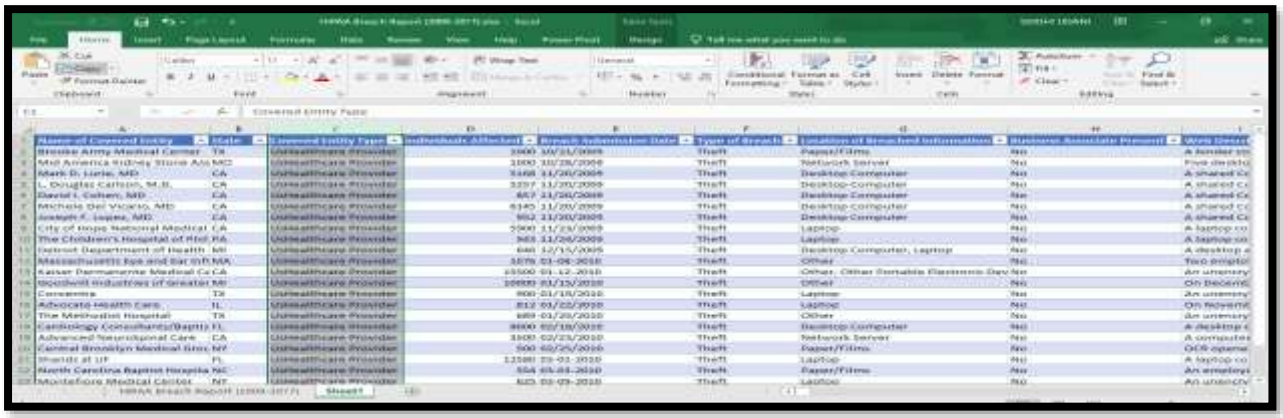


➔ After cleaning:

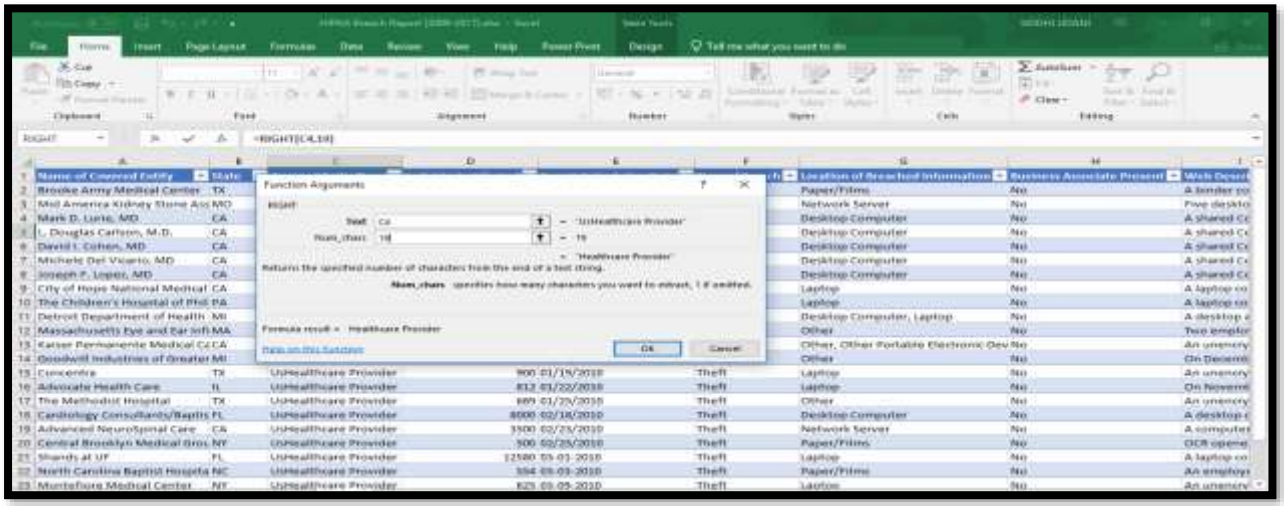


5. **ELIMINATING LEADING & TRAILING UNWANTED VALUES**: The dataset contained some columns with unwanted leading values, that were removed using the 'RIGHT' function from the 'TEXT' option under the 'FORMULA' toolbar, which returns the specified number of characters from the end of the text string, thereby cleaning it. Here, the 'UsHealthcare Provider' value was changed to 'Healthcare Provider' for better understanding. The screenshot represents this task for one specific cell at a time, however, the other cells too were cleaned using the same method.

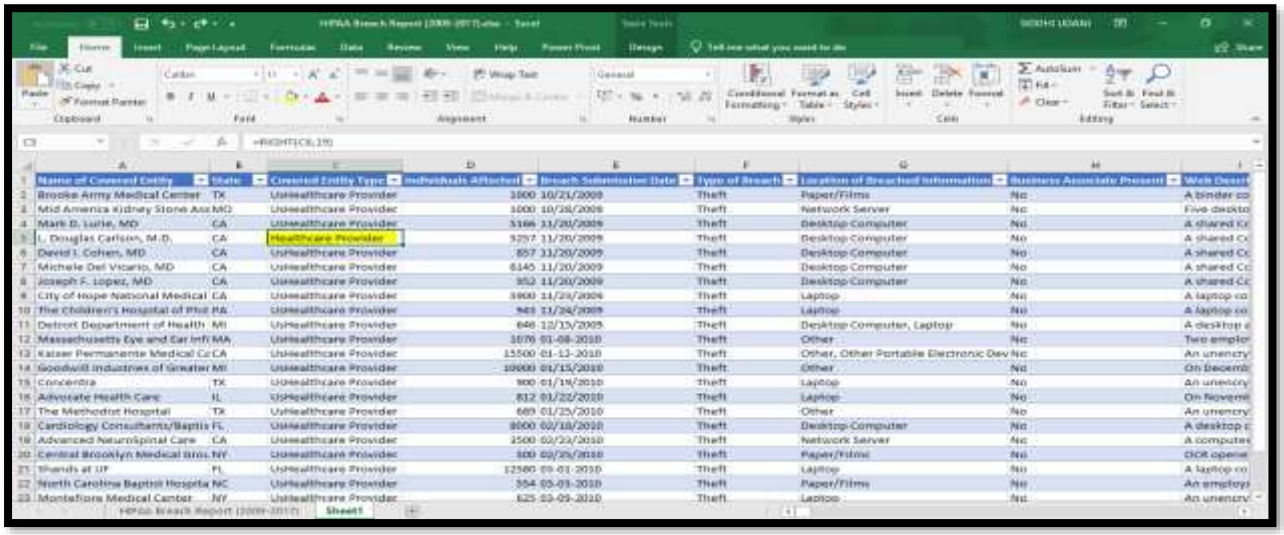
➔ Before cleaning:



➔ Cleaning step:



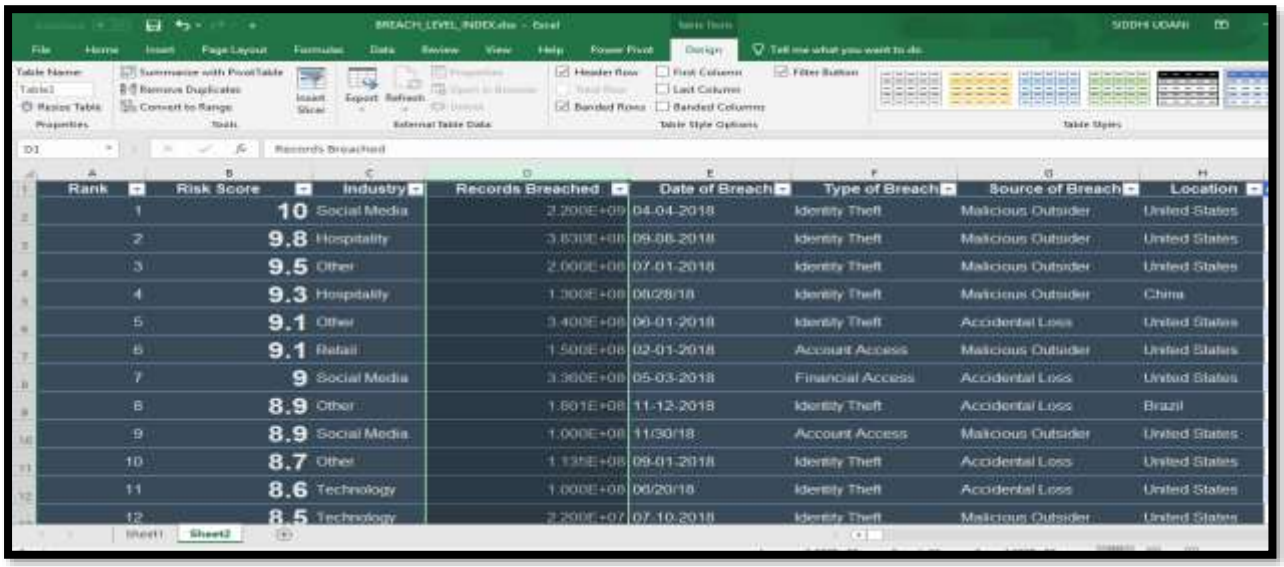
➔ After cleaning:



6. Converting numbers stored as 'SCIENTIFIC' into 'NUMBERS': In the Dataset, under the 'Records Breached' column, the column contained numeric values in the 'SCIENTIFIC' format, that were cleaned into normal 'NUMERIC' format, using the 'NUMBER' field from the 'FORMAT CELL' option, for better representation and

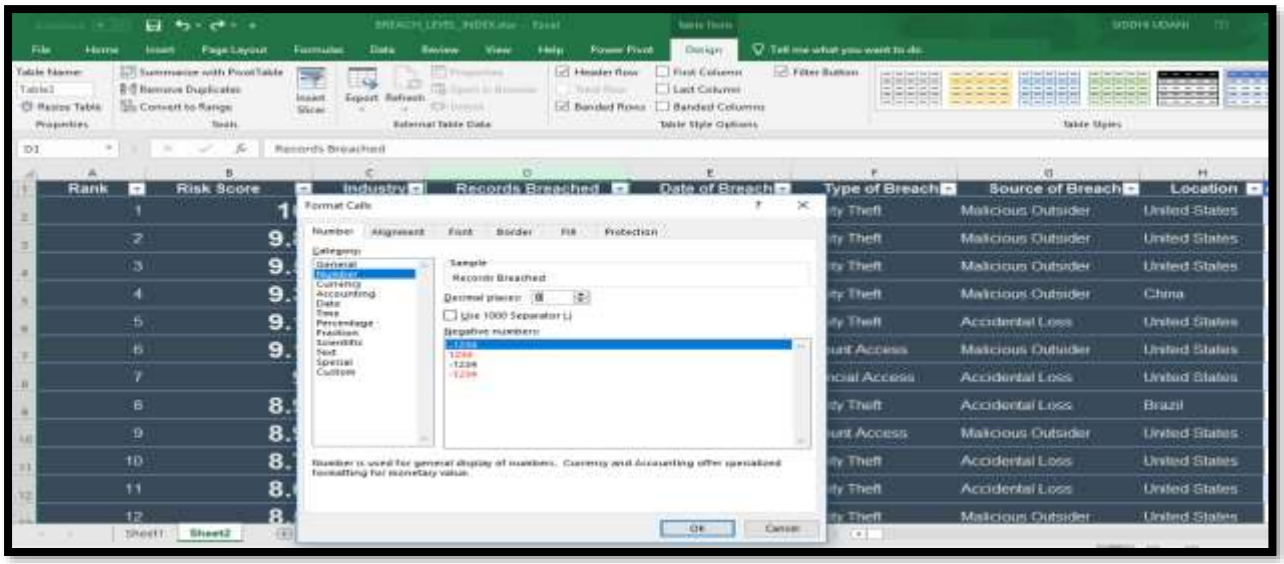
understanding, since the scientific format is difficult to interpret and not suitable for visualizations.

➔ Before cleaning:

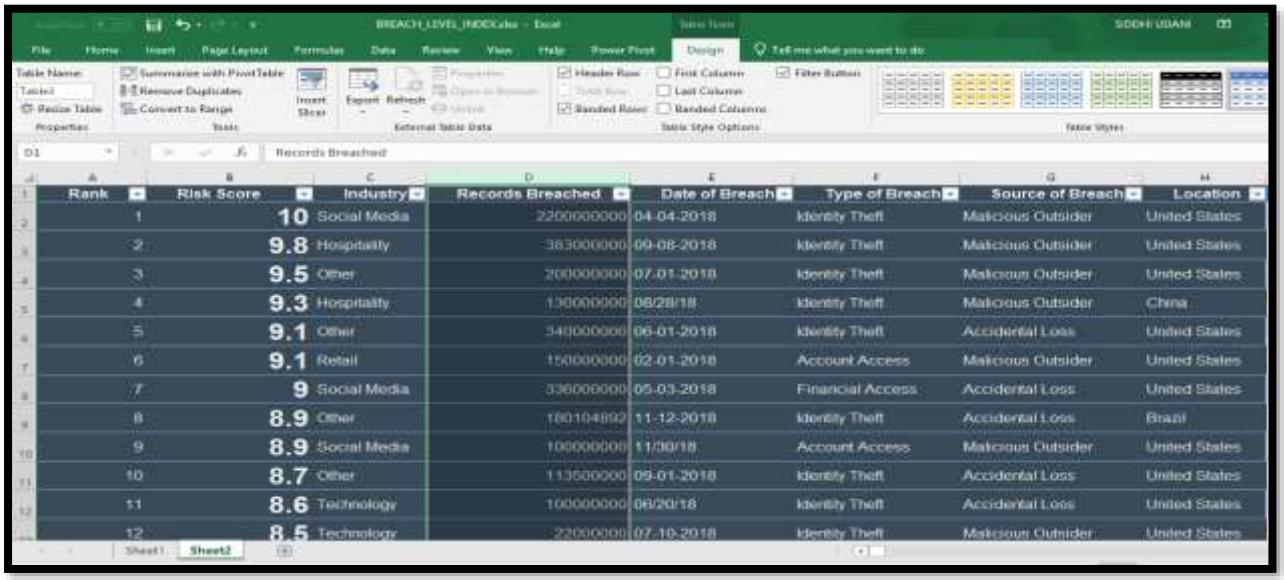


Rank	Risk Score	Industry	Records Breached	Date of Breach	Type of Breach	Source of Breach	Location
1	10	Social Media	2.200E+09	04-04-2018	Identity Theft	Malicious Outsider	United States
2	9.8	Hospitality	3.830E+08	09-08-2018	Identity Theft	Malicious Outsider	United States
3	9.5	Other	2.000E+08	07-01-2018	Identity Theft	Malicious Outsider	United States
4	9.3	Hospitality	1.300E+08	06/28/18	Identity Theft	Malicious Outsider	China
5	9.1	Other	3.400E+08	06-01-2018	Identity Theft	Accidental Loss	United States
6	9.1	Retail	1.500E+08	02-01-2018	Account Access	Malicious Outsider	United States
7	9	Social Media	3.300E+08	05-03-2018	Financial Access	Accidental Loss	United States
8	8.9	Other	1.801E+08	11-12-2018	Identity Theft	Accidental Loss	Brazil
9	8.9	Social Media	1.000E+08	11/30/18	Account Access	Malicious Outsider	United States
10	8.7	Other	1.135E+08	09-01-2018	Identity Theft	Accidental Loss	United States
11	8.6	Technology	1.000E+08	06/20/18	Identity Theft	Accidental Loss	United States
12	8.5	Technology	2.200E+07	07-10-2018	Identity Theft	Malicious Outsider	United States

➔ Cleaning step:



➔ After cleaning:

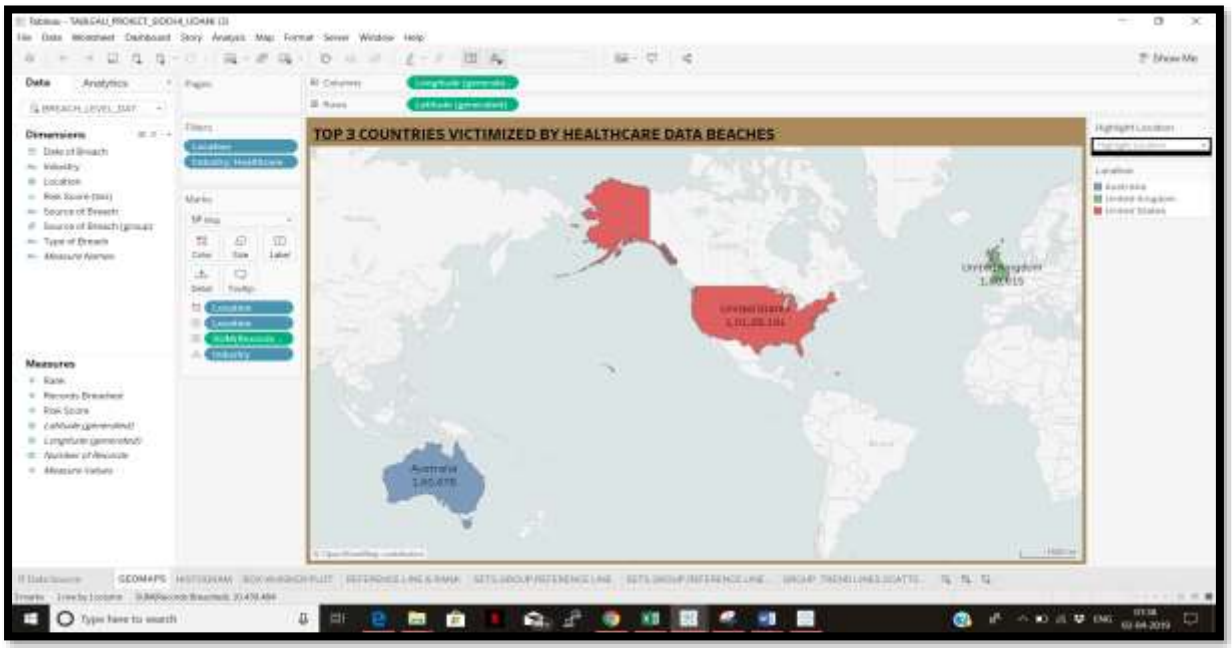


Rank	Risk Score	Industry	Records Breached	Date of Breach	Type of Breach	Source of Breach	Location
1	10	Social Media	2200000000	04-04-2018	Identity Theft	Malicious Outsider	United States
2	9.8	Hospitality	383000000	09-08-2018	Identity Theft	Malicious Outsider	United States
3	9.5	Other	200000000	07-01-2018	Identity Theft	Malicious Outsider	United States
4	9.3	Hospitality	130000000	08/28/18	Identity Theft	Malicious Outsider	China
5	9.1	Other	340000000	06-01-2018	Identity Theft	Accidental Loss	United States
6	9.1	Retail	150000000	02-01-2018	Account Access	Malicious Outsider	United States
7	9	Social Media	336000000	05-03-2018	Financial Access	Accidental Loss	United States
8	8.9	Other	180104892	11-12-2018	Identity Theft	Accidental Loss	Brazil
9	8.9	Social Media	100000000	11/30/18	Account Access	Malicious Outsider	United States
10	8.7	Other	113500000	09-01-2018	Identity Theft	Accidental Loss	United States
11	8.6	Technology	100000000	08/20/18	Identity Theft	Accidental Loss	United States
12	8.5	Technology	22000000	07-10-2018	Identity Theft	Malicious Outsider	United States

C. DATA VISUALIZATIONS:

1. Determine the top 3 countries where the highest number of healthcare records are being breached?

DATA VISUAL:



VISUALS USED: GEOGRAPHIC MAPS

- Map based on longitude and latitude were generated. Color shows details about location.The marks are labelled by location. Details are shown for Healthcare Industry. The view is filtered on location and Industry.
- The 2017 Breach Level Index from Gemalto shows that data breaches are very much a growing threat for organizations. The number of records compromised is remarkable, Broken down by region, **North America** led the way in the number of both compromised records and security incidents. Following this, **Australia** stood second by being victimized due to healthcare breaches and the third last victim being **United Kingdom.**

LOCATION (COUNTRY)	SUM OF RECORDS BREACHED
1.UNITED STATES	1,01,28,191
2.AUSTRALIA	1,60,678
3.UNITED KINGDOM	1,50,615

2. What is the trend of Risk score for the Healthcare Industry?

DATA VISUAL:

VISUALS USED: HISTOGRAM

- The Risk score determined how seriously an organization was affected when a data breach occured. DATA BREACH RISK ASSESSMENT CALCULATOR Calculates your own risk score and breach severity using the Breach Level Index data.
- The visualization shows the trend of sum of Risk score for Risk score(bin). Color shows the sum of records breached. The dataset is filtered on Industry, which keeps healthcare.
- The histogram depicts that a minimum of **16** records were breached with a minimum risk score of **17.2** whereas the maximum number of **1,37,40,000** records were breached with a risk score of **51.0**. The risk score has been mentioned on the basis of the sum of the records.

RISK SCORE BIN	SUM OF RECORDS BREACHED	CUMULATIVE RISK SCORE
0	16	17.2
1	223	77.8
2	5,877	77.3
3	96,782	309.9
4	372,807	482.6
5	1,986,200	303.9
6	1,557,612	68.4
7	13,740,000	51.0

3. Which are the top 5 states where the highest number of individuals were affected on the basis of the most common types of breaches?

DATA VISUAL:

<u>VISUALS USED: BOX AND WHISKER PLOT</u>

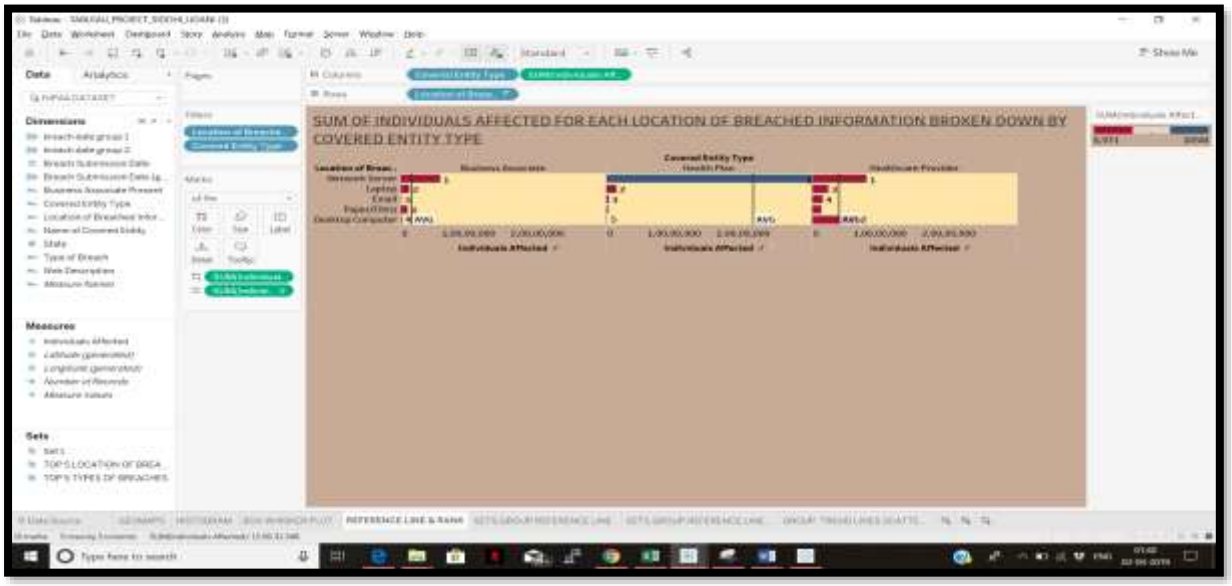
- The plot shows the Average of individuals affected for each state. Color shows details about type of breach. The view is filtered on state and type of breach. The state filter keeps California, Indiana, New York, Tennessee and Washington. The type of breach filter keeps Hacking/IT Incident, Improper disposal, Loss, Theft and Unauthorized access/disclosure.
- From this visual we understand that highest number of individuals were affected in the state of Tennessee due to 'Unauthorized access/disclosure' type of breach.

STATE	MAXIMUM OF AVG INDIVIDUALS AFFECTED	TYPE OF BREACH
CALIFORNIA	21510	**<u>THEFT</u>**

INDIANA	**15479**	**UNAUTHORIZED ACCESS/DISCLOSURE**
NEW YORK	**11679**	**THEFT**
TENNESSEE	**40357**	**UNAUTHORIZED ACCESS/DISCLOSURE**
WASHINGTON	**16217**	**UNAUTHORIZED ACCESS/DISCLOSURE**

4. How many individuals were affected for each location of breached information broken down by covered entity type?

DATA VISUAL:



VISUALS USED: REFERENCE LINE AND RANK

- The visualization shows the sum of individuals affected for each location of breached information broken down by covered entity type. Color shows sum of individuals

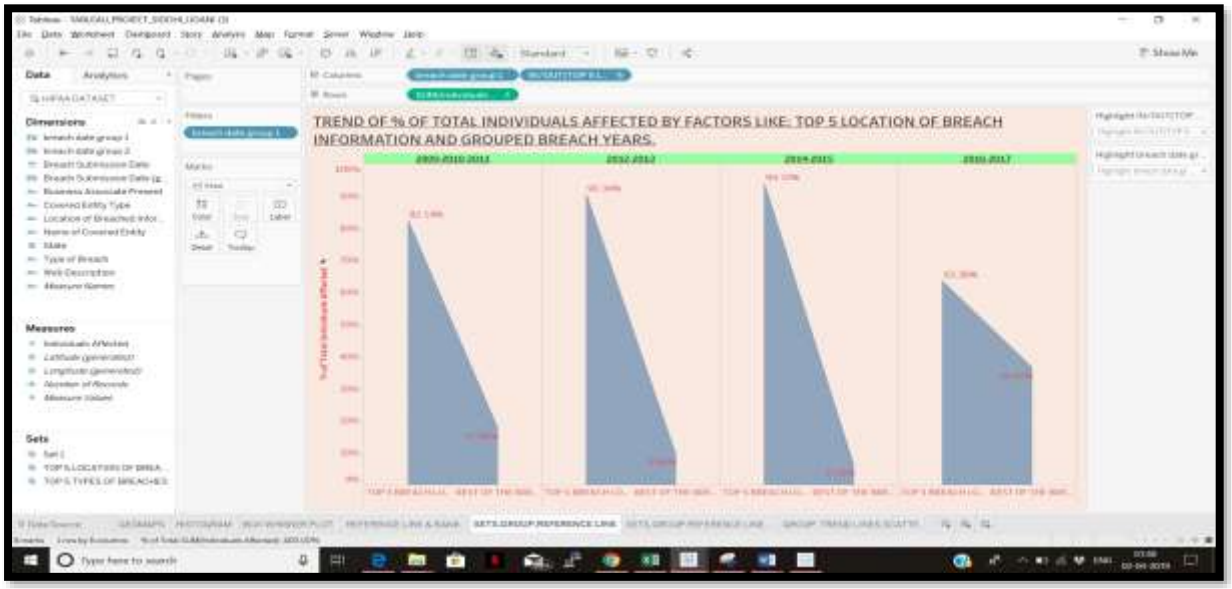
affected. The marks are labelled by Rank of individuals affected. The location of breached information filter has multiple members selected like Network server, Laptop, Email, Paper/films and Desktop computer.

- Covered entities are defined in the HIPAA rules as (1) Business associate, (2) health care plans, and (3) health care providers who electronically transmit any health information in connection with transactions for which HHS has adopted standards.
- The average reference line has been calculated for the covered entity types. This visual indicates that majority of the individuals having their covered entity types as: Business associate, health care plans and health care providers had most of their records breached due to 'NETWORK SERVER' glitches.

COVERED ENTITY TYPE	VALUE OF THE AVERAGE REFERENCE LINE	LOCATION OF BREACH (MAX INDIVIDUALS AFFECTED)	LOCATION OF BREACH (MIN INDIVIDUALS AFFECTED)
BUSINESS ASSOCIATE	1,633,086	NETWORK SERVER	EMAIL
HEALTH PLAN	21,506,832.2	NETWORK SERVER	DESKTOP COMPUTER
HEALTHCARE PROVIDER	3,986,351	NETWORK SERVER	PAPER/FILMS

5. What is the trend of % of total individuals affected by the top 5 location of data breaches information,yearwise?

DATA VISUAL:



<u>VISUALS USED: SETS AND GROUPS</u>

- The visual represents the % of total individuals affected for each top 5 location of breached information (in/out) broken down by breach date group 1. The view is filtered on breach date group 1,which keeps ‘2009-2010-2011’, ‘2012-2013’, ‘2014-2015’ and ‘2016-2017’. Percents are based on each row of each pane of the table.
- The sets for the ‘Location of breached information’ has been made as follows:

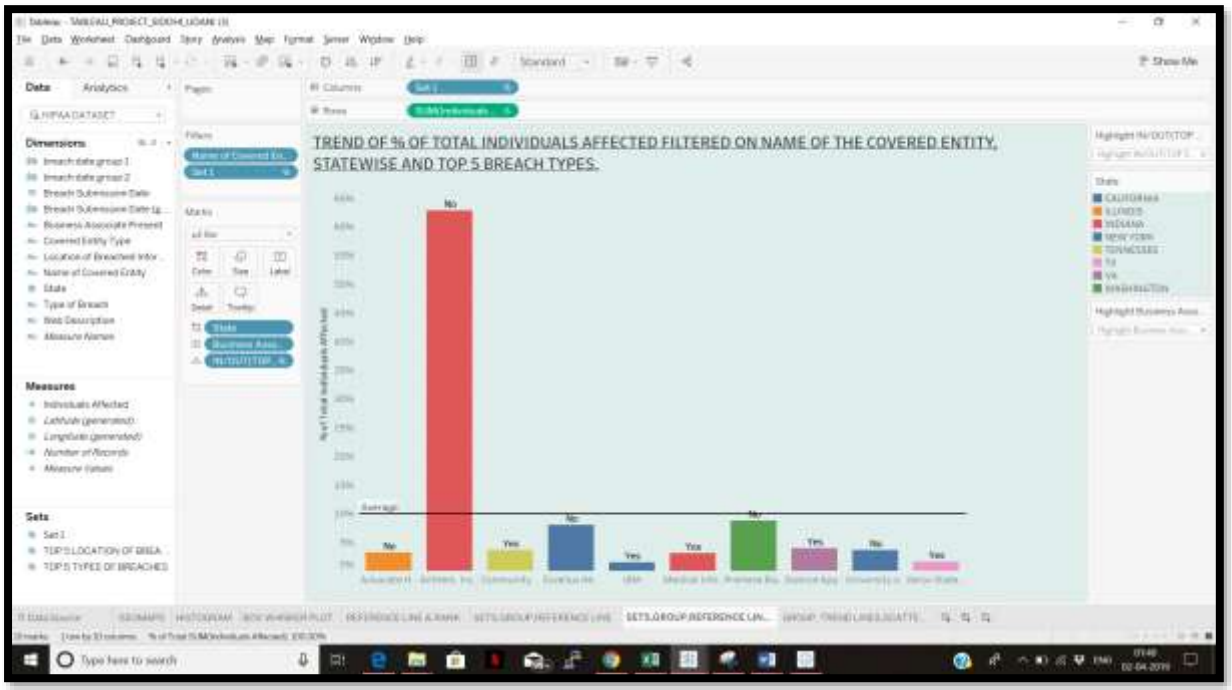
  -- Top 5 location of breached information includes (IN members): Desktop computer, Laptop, Network server, other and paper/films.

  --Rest of the breached location (OUT members): All the remaining locations of breached information.
- On interpreting the data we understand that the year from 2014-2015 recorded the highest number of data breached reports with a sudden decline in the year 2016 and 2017.

YEAR	% OF TOTAL INDIVIDUALS AFFECTED DUE TO TOP 5 LOCATION OF BREACHES	% OF TOTAL INDIVIDUALS AFFECTED DUE TO THE REMAINING LOCATION OF BREACHES
2009-2010-2011	82.14%	17.86%
2012-2013	90.34%	9.66%
2014-2015	93.72%	6.28%
2016-2017	63.38%	36.62%

6. What was the name of the covered entities in the top few states that suffered data breaches, indicating the presence/absence of business associate?

DATA VISUAL:



VISUALS USED: SETS, GROUPS AND REFERENCE LINE

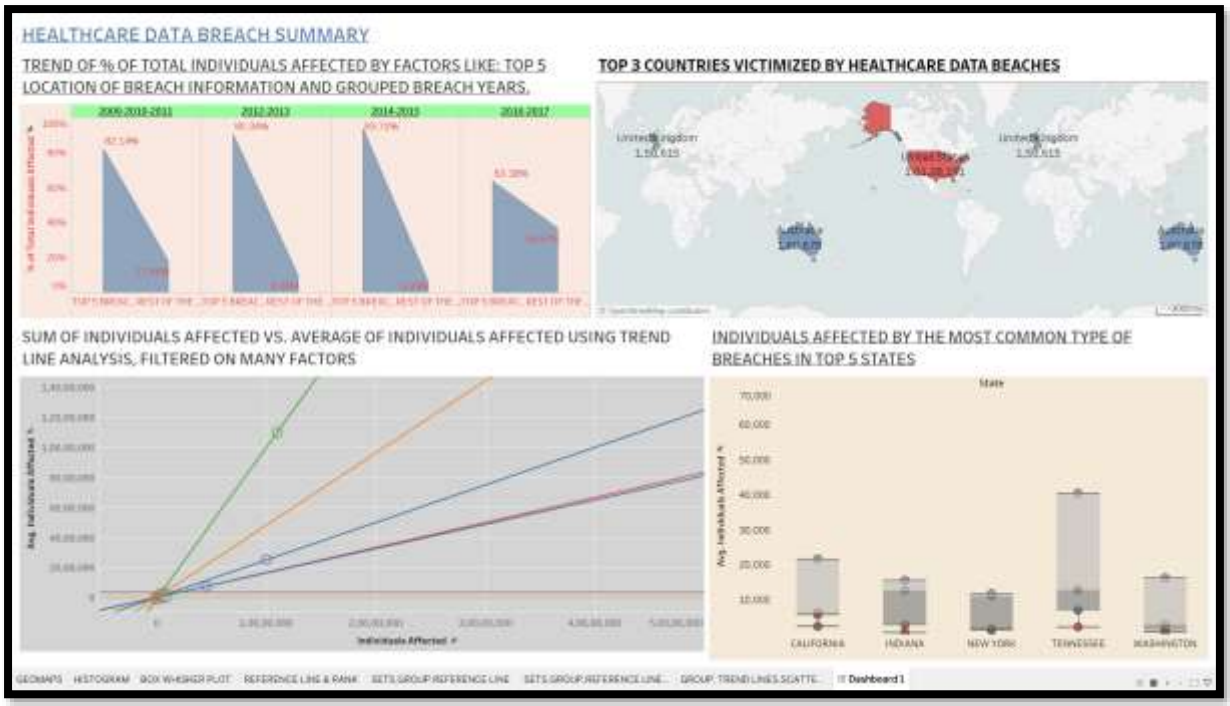
- The trend displays the % of total individuals affected for set 1, where the set 1 domain represents the top 10 names of the covered entities where the highest number of individuals where affected. The set 1 includes the members as viewed in the visual. Color shows details about state. The marks are labelled by business associate present. Details are shown for IN/OUT of TOP 5 TYPES OF BREACHES. Percents are based on each row of each pane of the table.
- The visual indicates that 62.77% of the total individuals were affected in the state of 'INDIANA' due to the top 5 types of breaches, where the NAME OF THE COVERED ENTITY is: Anthem, Inc. Affiliated covered entity, where there was NO business associate present.

NAME OF THE COVERED ENTITY	STATE	BUSINESS ASSOCIATE PRESENT	% OF TOTAL INDIVIDUALS AFFECTED
ADVOCATE HEALTH AND HOSPITALS MEDICAL GROUP	ILLINOIS	NO	3.21%
ANTHEM,INC AFFILIATED COVERED ENTITY	INDIANA	NO	62.77%

COMMUNITY HEALTH SYSTEMS PROFESSIONAL	TENNESSEE	YES	3.58%
EXCELLUS HEALTH PLAN	NEW YORK	NO	7.97%
IBM	NEW YORK	YES	1.51%
MEDICAL INFORMATICS ENGINEERING	INDIANA	YES	3.11%
PREMERA BLUE CROSS	WASHINGTON	NO	8.76%
SCIENCE APPLICATIONS INTERNATIONAL CORPORATION	VIRGINIA	YES	3.90%
UNIVERSITY OF CALIFORNIA,LOS ANGELES HEALTH	CALIFORNIA	NO	3.59%
XEROX STATE HEALTHCARE, LLC	TEXAS	YES	1.59%

D. <u>DASHBOARD:</u>

DASHBOARD DATA VISUAL:



The dashboard above shows four different and major elements for the discoveries in healthcare data breaches. The visual made using the geographic maps depicts the top 3 countries where the highest number of healthcare data breaches took place. The box and whisker plot determines the average of individuals affected in the top 5 states, being filtered on the most common type of data breaches. The linear trend line analysis using **scatter plot** is computed for the average of individuals affected. The area based graph visual, made using sets and groups, represents the year wise distribution of the percent of individuals affected. Thus, all of these key points were summarized using the dashboard feature of tableau.

VISUAL USED FOR ADDITIONAL VISUALIZATION ON DASHBOARD: SCATTER PLOT, TREND LINE ANALYSIS.

E. STORY TELLING:

HIPAA journal provides the most comprehensive coverage of the HIPAA breaches in addition to independent advice about HIPAA compliance and the best practices to adopt and avoid data breaches. The **Breach Level Index** indicates that the date breaches have been increasing in frequency and size over the last couple of years. This dataset on the other hand, explores the attributes of breaches like: number of records, date of breach, type of breach, source of breach, its location and the risk score.

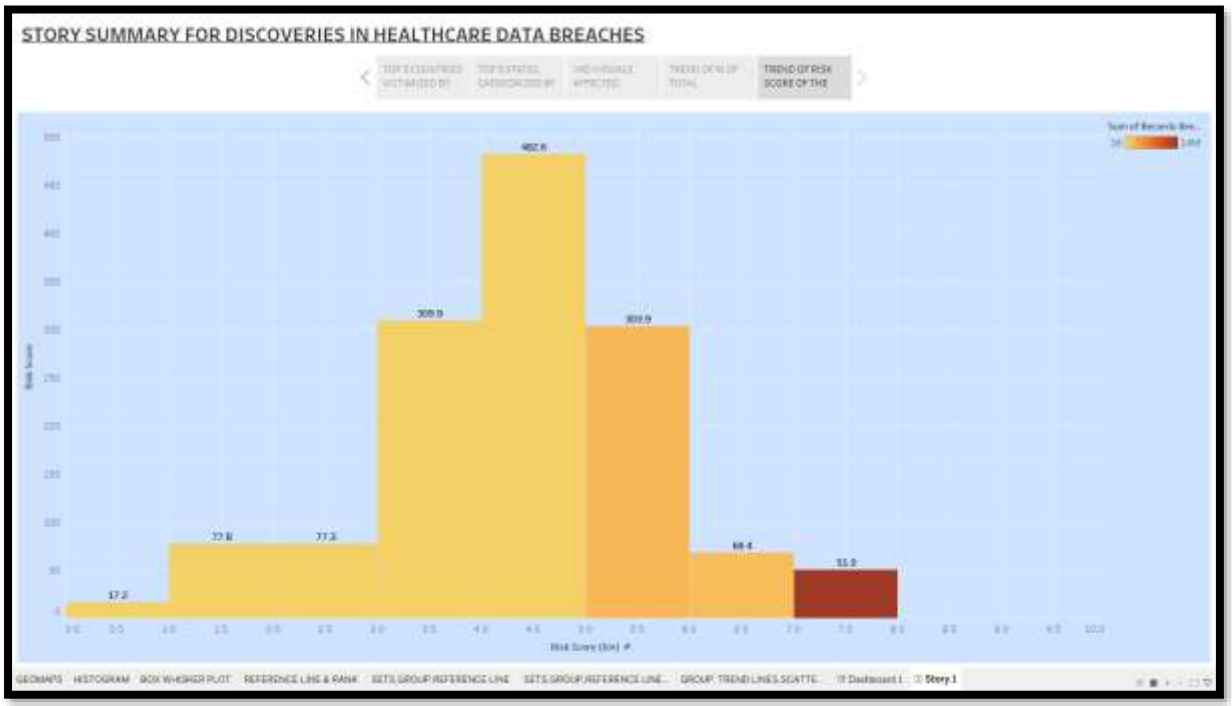


FIGURE: STORY TELLING CAPTION POINT

The threat landscape has continued to evolve throughout the year, with hackers ramping up targeted, sophisticated attacks. Ransomware continued to plague the healthcare sector, while phishing attacks and insider errors led to some of the biggest breaches in the recent years. However, resources and staffing gaps continues to be problematic. And hackers will continue to pummel the sector with targeted attacks through 2019 and beyond, globally.

Taking a look at the Breach level index report, out of all the countries in the world, the highest number of healthcare data breaches were targeted in **UNITED STATES**, followed by **AUSTRALIA** and **UNITED KINGDOM**. Since United States reported the highest number of breaches, it became important to determine the top particular states from where these breaches actually originated. After summarizing a visual over this question, I analysed that, **California**, **Indiana**, **New York, Tennessee** and **Washington** were the states which populated the count of healthcare data breaches.

Major security breaches in healthcare in the last few years that have resulted in the **exposure/theft** or due to **unauthorized access/disclosure of healthcare data records.** More than 41% of the population of the United States have had some of their protected health information exposed as a result of those breaches, which have been occurring at a rate of almost one a day over the past three years[1].

There has been a downward trend in the number of **theft/loss incidents** over the past three years as healthcare organizations have started encrypting records on portable electronic devices. However, improper disposal incidents have risen year over year as have hacking incidents. In **2017**, **hacking/IT incidents** were the main cause of healthcare data breaches[4] .

It seems that every day another hospital is in the news as the victim of a data breach. The routine is familiar - individuals receive notification by (e)mail of the breach due to **unauthorized access/disclosure**, **loss of information** to outsider sources, **hacking/IT Theft**, **improper disposal** of data. (One might wonder - Is there even anyone left who isn't being monitored?). The top 5 states in USA mentioned in the above explanation were determined on the basis of these types of breaches.

According to the Ponemon Institute and Verizon Data Breach Investigations Report[5], the health industry experiences more data breaches than any other sector. There may be

some potential for bias in this claim, due to the well-defined, legally mandated reporting requirements of the **Health Insurance Portability and Accountability Act (HIPPA),** which makes it more likely healthcare breaches will be reported compared to breaches in other sectors. The HIPAA dataset has various covered entity types who electronically transmit health information in connection with any transaction for which HHS(HEALTH AND HUMAN SERVICES) has adopted a standard, out of which the top 3 covered entities were: **BUSINESS ASSOCIATE**, **HEALTH PLAN**, **HEALTHCARE PROVIDER**. All of these breaches occur at a certain technical location, the **top 5 locations** being: **Network server**, **Laptop, Email**, **Paper/Films** and **Desktop computer**.

The visual indicated that healthcare data records were breached highest while on the Network server, amongst all the 3 covered entity types, in particular, which was the leading factor behind several breaches where data records were disclosed because organizations didn't take proper action to secure their cloud-based assets. That was the case especially with instances where companies violated users' privacy.

As per the previous analysis on the location of the breach, when we further dive deep in these records being breached yearwise, we determine that the year **2014** and **2015**(represented as 2014-2015), recorded the highest number of healthcare data breaches, as high upto **93.72%** in the top 5 location as mentioned above. More healthcare security breaches are being reported than at any other time since HIPAA required covered entities to disclose data breaches, although the number of individuals affected by healthcare data breaches has been declining year-over year for the past three years, and the visual exactly represents this, stating a decline of healthcare data breaches, as low as **63.38%** in the year **2016** and **2017.**

Breaches are widely observed in the healthcare sector and can be caused by many different types of incidents, including credential-stealing malware, an insider who either purposefully or accidentally discloses patient data, or lost laptops or other devices. Personal Health Information (PHI) is more valuable on the black market than credit card credentials or regular Personally Identifiable Information (PII). Therefore, there is a higher incentive for cyber criminals to target medical databases, so they can sell the PHI or use it for their own personal gain [2].

Each of the healthcare data breaches that take place are recorded with a certain level of risk associated with it, which indicates how severe the breach was and how badly it affected the organization as well as the individuals associated to it. From our histogram visual we depicted that the highest number of breach records, that is **1,37,40,000,** were affected with a breach score of **51.0,** which is really high. As per the Breach Level Index report for the year 2018, from an industry perspective, **Healthcare** companies experienced the greatest amount of security events in the years 2017 and 2018. Medical organizations were at the top of the list, though the number of incidents occurred was higher.[6].

These trends regarding data breaches look grim, but experts are working on ways to stop these breaches. The health care industry is comparatively unprepared when it comes to data security. Confronting the problem involves not only <u>understanding the threat</u>, but being proactive with combating it, which means not only solving old problems but racing to protect against new ones. It seems as though not a day goes by without a headline screaming that some organisation has experienced a data breach, putting the business – and its customers and partners – at risk. To keep your own organisation out of the news, it's important to understand the most common causes of data breaches and what you can do to mitigate the threats they present[3].

F. References:

1. Security Breaches in Healthcare in the Last Three Years. (2018, March 30). Retrieved March 9, 2019, from https://www.hipaajournal.com/security-breaches-in-healthcare-in-the-last-three-years/
2. The biggest healthcare data breaches of 2018 (so far). (2018, October 26). Retrieved February/March, 2019, from https://www.healthcareitnews.com/projects/biggest-healthcare-data-breaches-2018-so-far
3. Data Breaches: In the Healthcare Sector. (2017, April 24). Retrieved March 9, 2019, from https://www.cisecurity.org/blog/data-breaches-in-the-healthcare-sector/
4. 8 Most Common Causes of Data Breach. (2018, October 08). Retrieved March 9, 2019, from https://www.sutcliffeinsurance.co.uk/news/8-most-common-causes-of-data-breach/
5. Roiter, N. (2012, March 23). Ponemon, Verizon data breach cost, investigations reports show the way to actionable security intelligence. Retrieved March 9, 2019, from: https://www.corero.com/blog/99-ponemon-verizon-data-breach-cost-investigations-reports-show-the-way-to-actionable-security-intelligence.html
6. Gemalto. (2018, March 1). Data Breach Reports and Other Resources. Retrieved March 9, 2019, from https://breachlevelindex.com/data-breach-library