DATA BREACH ANALYSIS AMONG DIFFERENT ORGANIZATION BECAUSE DATA SECURITY IS NO LONGER AN UNDERDOG

CIS 5270: BUSINESS INTELLIGENCE

Spring 2019

By: Maitri Shah

Submitted to:

Dr. Shilpa Balan

A) Dataset URL

https://www.privacyrights.org/data-breaches

why privacyrights.org?

Records Breached: 11,583,442,497 from 9,094 DATA BREACHES made public since 2005.

There are so many researches going on data breach as it is a sensitive topic for which actions and some safety steps are required to protect the data. Data-risk analytics ought to be applied to every type of root-cause hacking stress. Data breach dataset which is collected from privacyrights comes from many different organizations and there are so many sources which are working to help on data breaches. This happens to a great extinct in USA (includes city, states along with longitude and latitude) from the year 2005 – 2019. This allow researchers who are working on data breach issue to know the details about its happenings and to get the insights which can helpful.

The organization that are part of dataset are:

1. BSF: Businesses-Financial and Insurance Services
2. BSO: Businesses – Other
3. BSR: Businesses-Retail/Merchant - Including Online Retail
4. EDU: Educational Institutions
5. GOV: Government & Military
6. MED: Healthcare, Medical Providers & Medical Insurance Services
7. NGO: Nonprofits

(B) Data Cleaning:

1. Standardization (Date Format):

The date format which needs to be in the standard form of was not available same so cleaned the date column for every row:

Before

Date Made Public	Company	Location
7-Oct-15	LoopPay	Woburn, Massachusetts
Sep 1 2016	New York State Psychi	New York, New York
8-Jan-17	E-Sports Entertainmer	Cologne, berlin
Jan 10 2017	Legal Aid Society of O	Santa Ana, California
Jan 10 2017	MetroPlus Health Plar	New York, New York
9-Jan-17	Kevin Harrington, CPA	Rancho Cordova, California
Jan 10 2017	SwimOutlet.com	Campbell
13-Jan-17	Children's Hospital of	Los Angeles
19-Jan-17	CoPilot Provider Servi	New Hyde Park

After

Date Made Public	Company	City	State
7-Oct-15	LoopPay	Woburn	Massachusetts
1-Sep-16	New York State Ps	New York	New York
8-Jan-17	E-Sports Entertain	Cologne	Berlin
10-Jan-17	Legal Aid Society (	Santa Ana	California
3-Jan-17	MetroPlus Health	New York	New York
9-Jan-17	Kevin Harrington,	Rancho Cordov	California
12-Jan-17	SwimOutlet.com	Campbell	California
13-Jan-17	Children's Hospita	Los Angeles	California
19-Jan-17	CoPilot Provider S	New Hyde Park	New York

2. Cleaned column Location: There was city as well as state mentioned in the column location which needs to be cleaned because insights as per the state and city is required for visualization.

Before

C	D	E	F
Location	State	Type of breach	Type of organization
Woburn, Massachusetts	Massachusetts	HACK	BSO
New York, New York	New York	HACK	GOV
Cologne, berlin	Berlin	HACK	BSO
Santa Ana, California	California	DISC	NGO
New York, New York	New York	DISC	MED
Rancho Cordova, California	California	HACK	BSF
Campbell	California	HACK	BSO
Los Angeles	California	PORT	MED

After

C	D	E	F
City	State	Type of breach	Type of organization
Woburn	Massachusetts	HACK	BSO
New York	New York	HACK	GOV
Cologne	Berlin	HACK	BSO
Santa Ana	California	DISC	NGO
New York	New York	DISC	MED
Rancho Cordova	California	HACK	BSF
Campbell	California	HACK	BSO
Los Angeles	California	PORT	MED

3. Removing special character:

   After the location has been formatted as city and state individual, it was found that it had many special characters. So, removed special characters

Before

Type of breach	Type of organization	Total Records
HACK	BSO	0
PORT	MED	1,586
HACK	BSF	7
HACK@	BSO	0
HACK!	BSO	0
PHYS@	MED	2,953
INSD*	MED	600
DISC	MED	703

After

Type of breach	Type of organization	Total Records
HACK	BSO	0
PORT	MED	1,586
HACK	BSF	7
HACK	BSO	0
HACK	BSO	0
PHYS	MED	2,953
INSD	MED	600
DISC	MED	703

4. Corrupt data:

   There were many corrupted data in the date column which had special characters, so

cleaned the date column

Before

Date Made Public	Company	Location
15-Feb-17	Operating Engineers L	Alameda
######	Jeffrey D. Rice	Zanesville
17-Feb-17	Deboer Income Tax	Big Bear Lake
######	Intex Recreation Corp	Long Beach
22-Feb-17	JoFit	Warminster
######	Catalina Post-Acute ar	Tucson
$$$$$$	Dignity Health St. Jose	Phoenix
17-Feb-17	Group Health Incorpor	New York

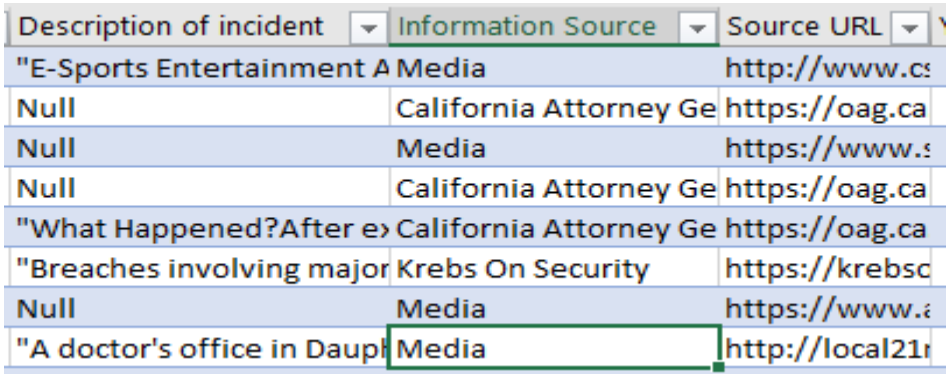
After

Date Made Public	Company	Location
15-Feb-17	Operating Engineers L	Alameda
2-Feb-17	Jeffrey D. Rice	Zanesville
17-Feb-17	Deboer Income Tax	Big Bear Lake
20-Feb-17	Intex Recreation Corp	Long Beach
22-Feb-17	JoFit	Warminster
23-Feb-17	Catalina Post-Acute ar	Tucson
23-Feb-17	Dignity Health St. Jose	Phoenix
17-Feb-17	Group Health Incorpor	New York

5. Removing NULL values:

   Source URL from which the data was found had many NULL values mentioned. So, deleted and cleaned all those data.

   Before



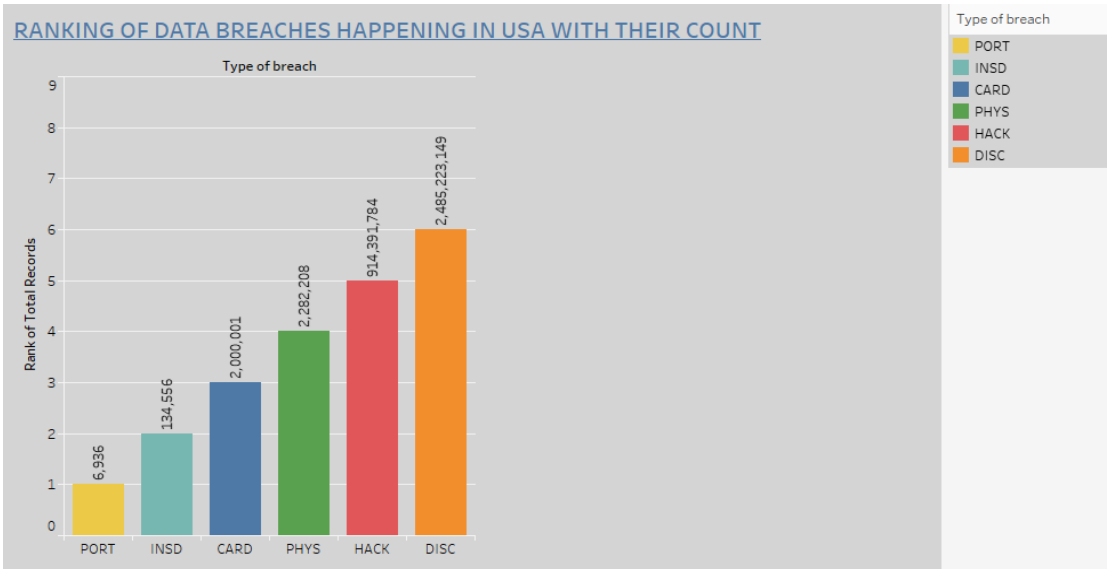
Description of incident	Information Source	Source URL
"E-Sports Entertainment A	Media	http://www.cs
Null	California Attorney Ge	https://oag.ca
Null	Media	https://www.s
Null	California Attorney Ge	https://oag.ca
"What Happened?After e>	California Attorney Ge	https://oag.ca
"Breaches involving major	Krebs On Security	https://krebsc
Null	Media	https://www.a
"A doctor's office in Daupl	Media	http://local21r

After

Description of incident	Information Source	Source URL
"US newspaper and media	Media	https://www.v
"Recently, many people re	Media	https://www.v
"A local car wash isn't the	Media	http://www.b
"McDavid, Inc. ("McDavid"	California Attorney Ge	https://oag.ca
"We are writing to inform	Vermont Attorney Ger	http://ago.ver
"On February 27, 2017 we	Vermont Attorney Ger	http://ago.ver
"This follows on the inforr	Vermont Attorney Ger	http://ago.ver
"Hill Country has recently	Vermont Attorney Ger	http://ago.ver

C) Data Visualization

Question 1: Ranking has been assigned to Data Breaches happening in USA with their total count



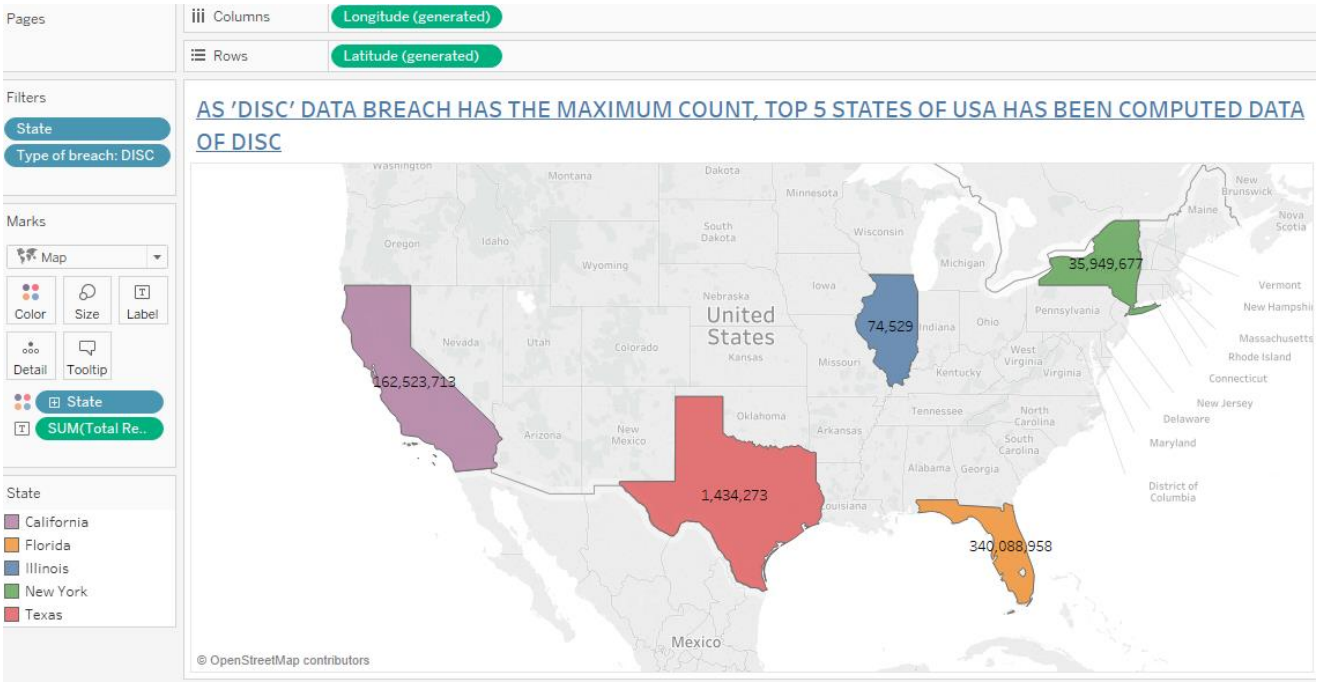
[Visualization used: Bar chart(Ranks assigned to each Data Breach)]

There are 6 types of breaches (1. PORT: Portable Devices, 2. INSD: Insider, 3.CARD: Payment card Fraud, 4. PHYS: Physical Loss, 5. HACK: Hacking, 6. DISC: Unintended Disclosure) observed in USA from the year 2017 and 2018.

From the BAR chart above which shows the ranks and storted as per the total number of record found. Ranking has been assigned to each of the data breach as per their total records. Calculated

ranking of individual breach among which DISC has the maximum count which is 2,485,223,149. Total count has been displayed to every data breach and arranged in ascending order.

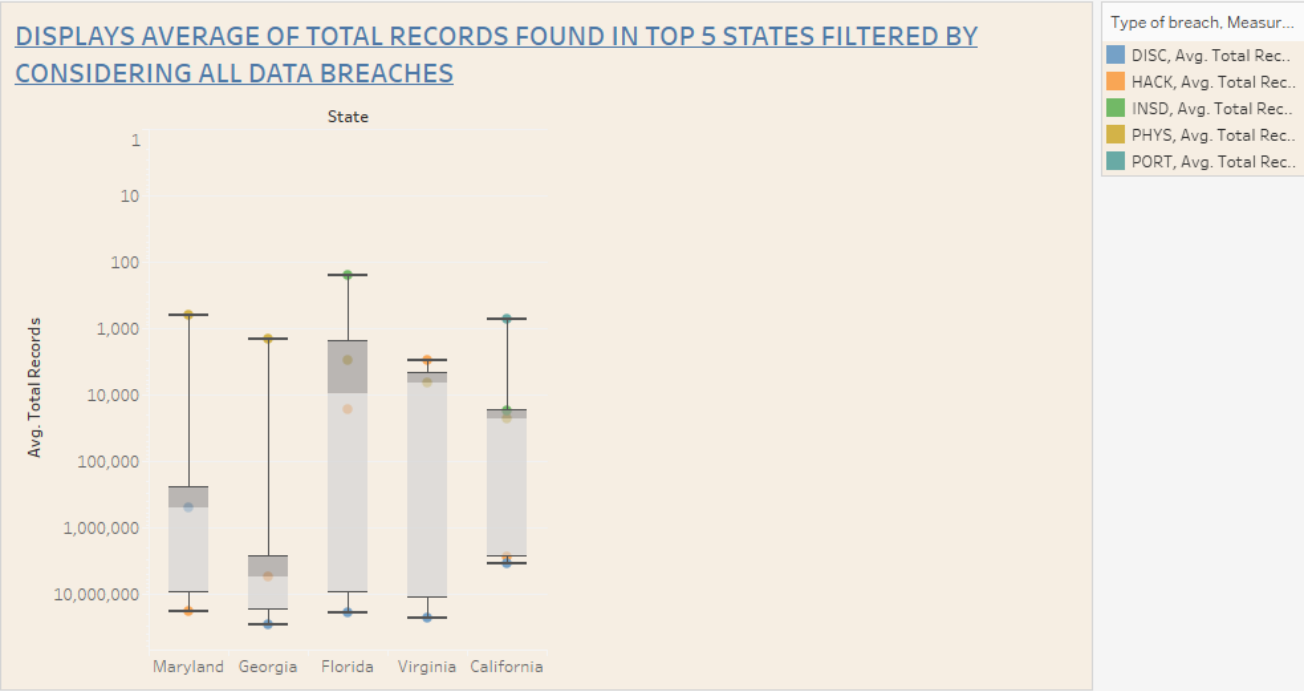
Question 2: Top 5 states as per the Data Breach DISC where maximum record is found



[**Visualization used: Geographic Maps**]

As per the previous visualization, it is found that the data breach 'DISC' has maximum count in USA and so more insight about the same needs to be computed. Filters is required to retrieve. One for the type of breach as only 'DISC' data breach is needed and another one is for the state which will display top 5 states along with their count in the map of USA. California, Florida, Illinois, New York, Texas are the top 5 states which has the total of 2,485,223,149.

Question 3: Displays average of total records found in Top 5 states filtered by considering all Data Breach

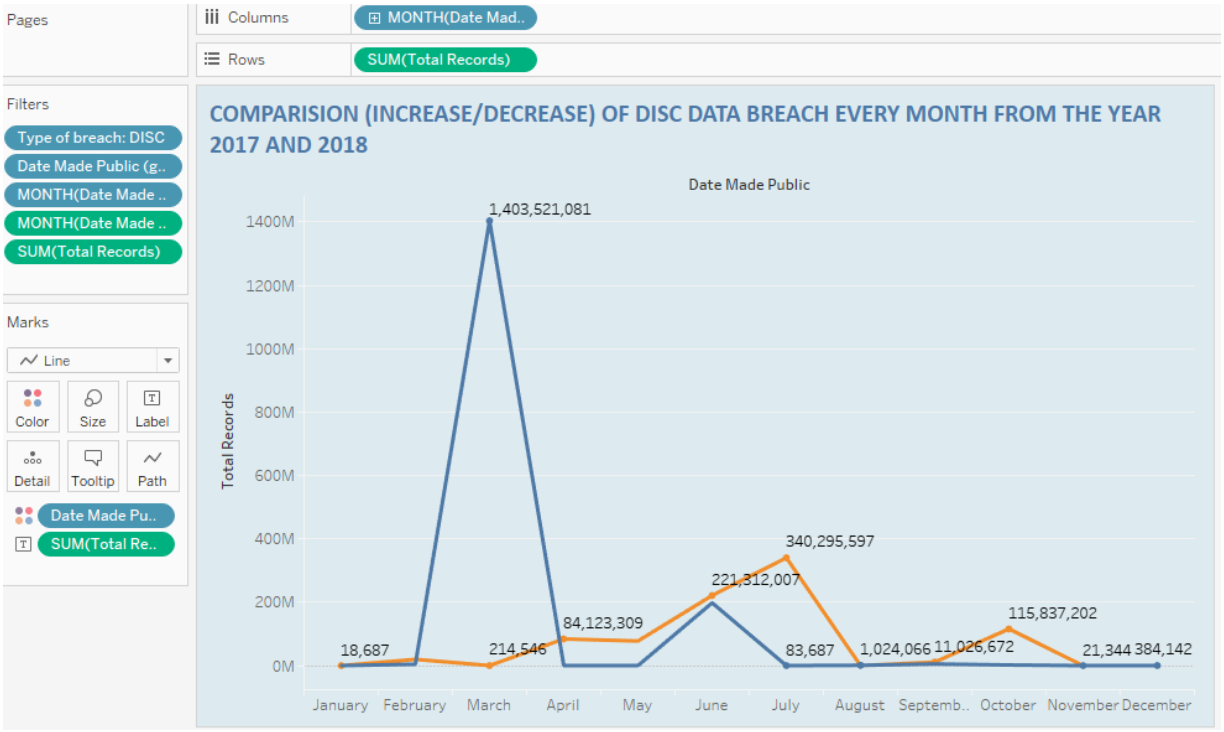


[Visualization Used: Box & Whiskers Plot]

Top 5 states called Maryland, Georgia, Florida, Virginia, California are observed in USA that has maximum data breach count. This is Box and Whiskers Plot that shows total data breach year 2017 and 2018.

From above plot we can say that for the PHYS data breach California has the maximum count, HACK has been maximum found in Maryland, DISC and INSC has been maximum found in California. Looking at the plot we can conclude that a lot of data breach has happened among which Maryland, Georgia, Florida, Virginia, California is the most affected.

Question 4: Comparison of breaches in 2017 and 2018 to more detail of its increase or decrease among top 5 states



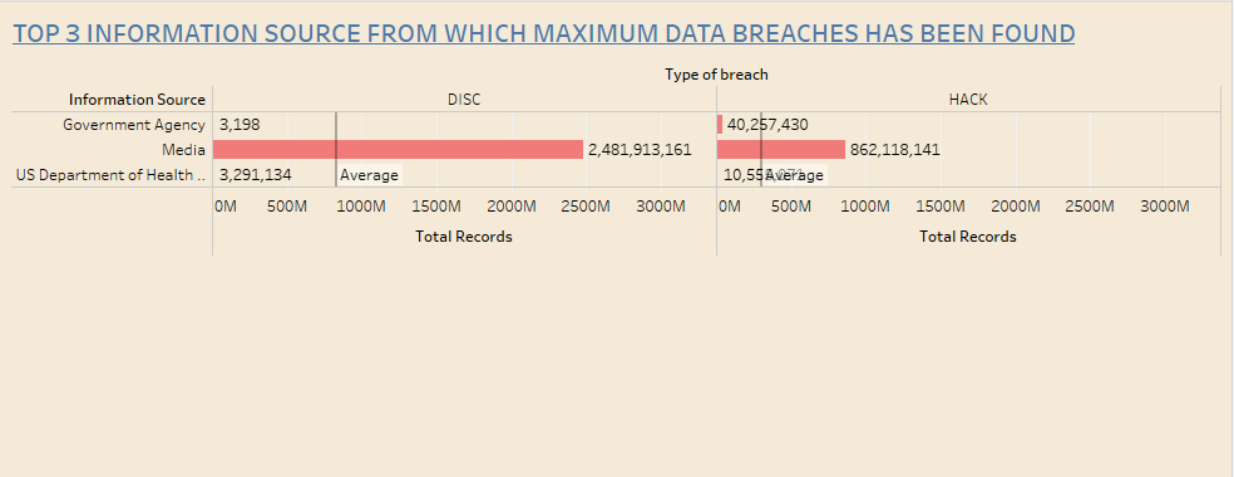
[**Visualization Used**: **Groups, line(discrete) chart**]

To get insights of Data BREACH "DISC" which was found as having maximum in previous visualizations. Individual Groups has been created for year 2017 and 2018 which can compare increase/decrease of data breach 'DISC' count every month (January to December)

As shown in the graph, it is found that in the month of March there is maximum count which is 1,403,521,081(year 2017) which decreased to 214,546 (year 2017).

It is visible from the chart that there is decrease in the month January, march, November and December. And for others there is increase but as we can see that in the month of March there was so much data record that overall count has been decreased from year 2017 to 2018.

Question 5: Information source from which maximum breaches are found which can help for further insights.

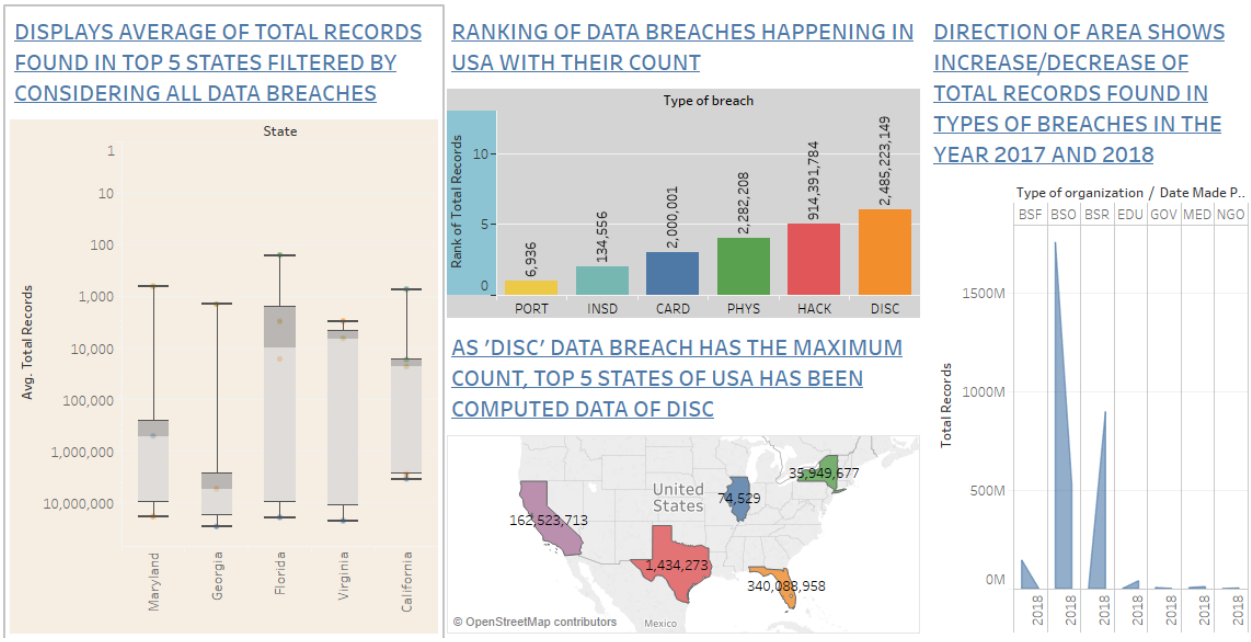


[**Visualization Used: Reference Line, Parameter**]

There are so many information sources from where the data is gathered. This source includes Media, Government agency, US Department of Health, California Attorney General,

Vermont Attorney General, Krebs On Security and many more. Among which top 3 sources are filtered and calculated along with the top 2 data breaches that is DISC and HACK. Reference lines has been used which shows the average for HACK as 304,312,214 as of total among which media is the first in the list with the total 862,118, 141. On the other hand, the max data breach DISC which has the average of 828,402,497.66 where again records are found in media 2,481,913,161.

(D) Dashboard



(E) Storytelling

Data Breach also called as Data leakage is an incident where information is stolen or taken from a system without the knowledge or authorization of the system's owner. A small company or large organization such as corporations and government agencies may suffer a data breach. The loss of sensitive information can lead to significant reputational damage and financial losses, and even can be detrimental to the long-term stability of an organization [1].

Stolen data may involve sensitive, proprietary, or confidential information such as credit card numbers, customer data, trade secrets, or matters of national security, employee/customer data, intellectual property, to medical records [3].

As per the article published on DZone Security, 2016 was a year of mega breaches. LinkedIn, MySpace, Yahoo and several other well-known online services had suffered

some sort of data breach. Billions of customers' records from these breaches have been traded on darknet marketplaces or privately between hackers which indicates that data breach is a serious issue and analysis for the same needs to be done [4].

First thing was found the top 5 states where data breaches are happening to the maximum which shown by the scatter plot. Affected states are Maryland, Georgia, Florida, Virginia and California.

Every year Data Breaches has decreased but it is not as per the needs and security has been a serious problem to all of these. So, dataset has been downloaded for data breaches and information is gathered for year 2017 and 2018. There are 6 types of breaches (1. PORT: Portable Devices, 2. INSD: Insider, 3.CARD: Payment card Fraud, 4. PHYS: Physical Loss, 5. HACK: Hacking, 6. DISC: Unintended Disclosure) observed in USA from the year 2017 and 2018.

Ranking has been calculated for individual breaches among which DISC has the maximum count which is 2,485,223,149. Hack on the 2nd highest ranking with count of 914,391,784, then comes PHYS, CARD, INSD and last is the PORT.

So further insights has been found for the data breach DISC as it has maximum count. Top 5 states where maximum DISC records are found are California (having maximum count of 162, 523,713), Florida, Illinios, New York and Texas(least count among top 5 states of 74529).

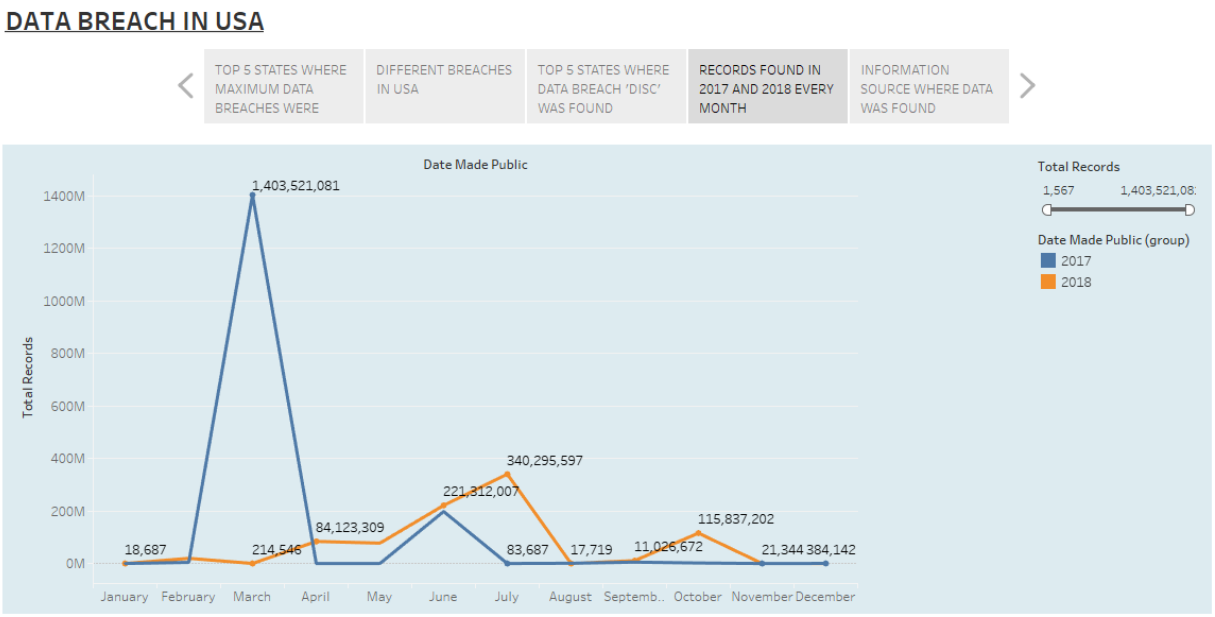
To get insights of Data BREACH "DISC" which was found as having maximum in previous visualizations, more insights for every month is calculated for the year to show the comparison of month by month in year 2017 and 2018.

Individual Groups has been created for year 2017 and 2018 which can compare

increase/decrease of data breach ‘DISC’ count every month (January to December)

As shown in the graph, it is found that in the month of March there is maximum count which is 1,403,521,081(year 2017) which decreased to 214,546 (year 2017).

It is visible from the chart that there is decrease in the month January, march, November and December. And for others there is increase but as we can see that in the month of March there was so much data record that overall count has been decreased from year 2017 to 2018.



There are so many information sources from where the data is gathered. This source includes Media, Government agency, US Department of Health, California Attorney General,

Vermont Attorney General, Krebs On Security and many more. Among which top 3 sources are filtered and calculated along with the top 2 data breaches that is DISC and HACK. Reference lines has been used which shows the average for HACK as 304,312,214 as of total among which media is the first in the list with the total 862,118, 141.

On the other hand, the max data breach DISC which has the average of 828,402,497.66 where again records are found in media 2,481,913,161.

According to the sources for information in year 2016 Data Breaches Incidents:

According to the non-profit organization ID Theft Resource Centre, in 2016 there were more than 1,000 data breaches in the US alone, a 40% increase from 2015. In this document, we only use a small sample from this data. We've chosen the incidents to use in these statistics based on the following criteria:

➢ The affected organization is well known.

➢ The data breach affected a wide range of people.

➢ The method of the attack that lead to the data breach is known (except for the Mega breaches).

➢ The date of the breach notification or when the data was leaked is 2016.

➢ Let's dive into the numbers.

Overall there were 61 breaches that matched the above-mentioned criteria. Through these breaches a total of 2.7 billion (2,716,114,000) confidential records were leaked [4].

References:

1. Cheng, L., Liu, F., Danfeng, Daphne, & Yao. (2017, June 09). Enterprise data breach: Causes, challenges, prevention, and future directions. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1211

2. Data Breaches. (n.d.). Retrieved from

https://www.privacyrights.org/data-breaches

3. Data Breach. (n.d.). Retrieved from

https://www.trendmicro.com/vinfo/us/security/definition/data-breach

4. Morgenroth, S. (2017, September 16). Analysis of Data Breaches - DZone Security. Retrieved from https://dzone.com/articles/analysis-of-data-breaches-2016