School of Computer Science / Faculty of Engineering & IT

ASSIGNMENT/PROJECT COVER SHEET – Group Assignment

Unit of Study: **ISYS5050 Knowledge Management System**

Assignment Name: ISYS5050 Final Project

DECLARATION

We the undersigned declare that we have read and understood the *University of Sydney Academic Dishonesty and Plagiarism in Coursework Policy*, an, and except where specifically acknowledged, the work contained in this assignment/project is our own work, and has not been copied from other sources or been previously submitted for award or assessment.

We understand that failure to comply with the *Academic Dishonesty and Plagiarism in Coursework Policy* can lead to severe penalties as outlined under Chapter 8 of the *University of Sydney By-Law 1999* (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.

We realize that we may be asked to identify those portions of the work contributed by each of us and required to demonstrate our individual knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

SIT Building, J12
The University of Sydney NSW 2006 Australia
T +61 2 9351 3423 F +61 2 9351 3838 E sit.info@sydney.edu.au
sydney.edu.au/it
ABN 15 211 513 464 CRICOS 00026A

Table of Contents

Introduction

Over time the breaches are getting more severe, more frequent, and more intense. Many individuals/organizations are getting compromised by their attacks. Therefore, to analyze the severity, intensity & frequency we have used the Cybersecurity data breaches dataset, refers to the security breaches in the United States from 2005 to 2018. It consists of the information about the breaches and the types of organizations/individuals that were impacted during this period.

To analyze the dataset, we have used visualization tools (as Tableau, OLAP) and for the refinement of messy and unwanted data, we have used Jupyter Notebook just to give better insights into the data/dataset.

Dataset: The cybersecurity dataset consists of 9016 records with 13 attributes. The attributes are as follows:

Type of Breaches: It consists of 8 different types of breaches

States: List of states in the US

Total Records: Records that are somehow compromised or impacted by the breaches

City: List of cities in each state of the US

Description of Incident: It describes an incident such as what kind of breach happened, who are comprised, and what they had lost

Year of Breach: Give the information of in which year the breach happened.

Types of Organization: what kind of organization gets impacted by the breach

There are many more attributes available such as Latitude, longitude, source URL

To analyze the data, we have segregated our tasks/reports into different sub-section.

a) **Data-preprocessing:** To support our claim we have done mathematical calculations and added a few more columns which we will describe later.

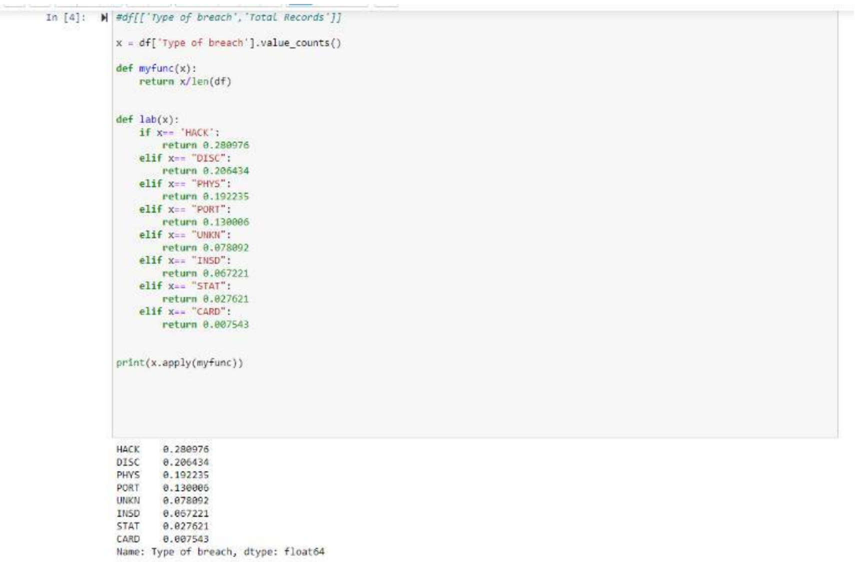
b) **Visualization:** To analyze the severity, frequency of occurrence, etc. over the years we have tried to answer the different questions by using different attributes and different visualizations.

c) **Our inference**

Data Pre-processing

Pre-processing was done in the analysis for data extraction and data transformation. We followed the below steps for pre-processing of data:

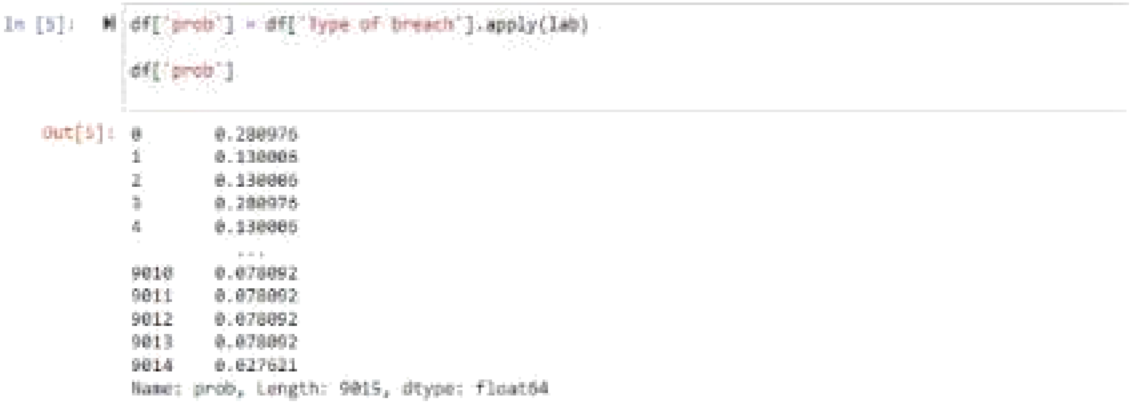
- The null values from the ‘Types of breaches’ were removed.
- To calculate the Severity of the breaches, two new columns were introduced ‘prob’ and ‘Severity’. Jupyter notebook software was used to calculate Severity.
- Prob: While calculating the severity, a new column was created called “prob” which basically gives the probability of each type of breach. The code to create this column has been shown below.



```
#df[['Type of breach','Total Records']]

x = df['Type of breach'].value_counts()

def myfunc(x):
    return x/len(df)


def lab(x):
    if x== 'HACK':
        return 0.280976
    elif x== "DISC":
        return 0.206434
    elif x== "PHYS":
        return 0.192235
    elif x== "PORT":
        return 0.130006
    elif x== "UNKN":
        return 0.078092
    elif x== "INSD":
        return 0.067221
    elif x== "STAT":
        return 0.027621
    elif x== "CARD":
        return 0.007543


print(x.apply(myfunc))
```

```
HACK    0.280976
DISC    0.206434
PHYS    0.192235
PORT    0.130006
UNKN    0.078092
INSD    0.067221
STAT    0.027621
CARD    0.007543
Name: Type of breach, dtype: float64
```

Figure 1: Python code to find the probability of occurrence of Breaches



```
In [5]:  df['prob'] = df['Type of breach'].apply(lab)

         df['prob']

Out[5]:  0       0.280976
         1       0.130006
         2       0.130006
         3       0.280976
         4       0.130006
                   ...
         9010    0.078092
         9011    0.078092
         9012    0.078092
         9013    0.078092
         9014    0.027621
         Name: prob, Length: 9015, dtype: float64
```

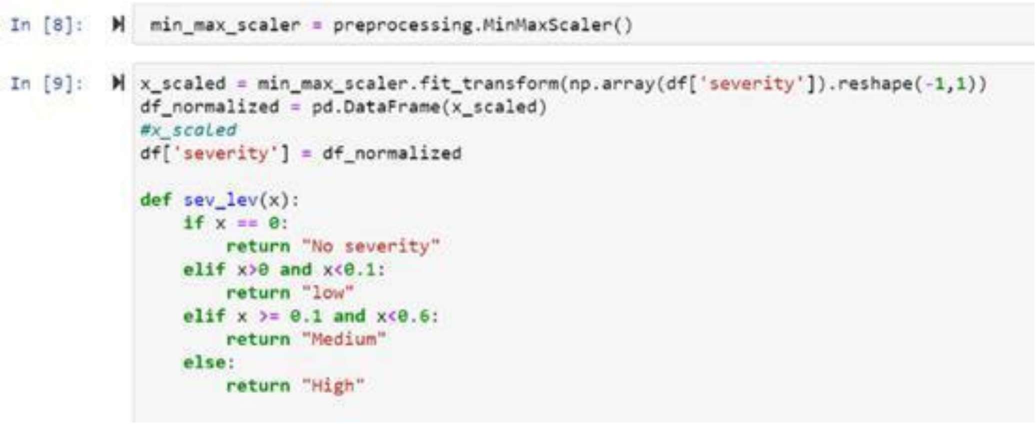
Figure 2: Python code to add the 'prob' column

- Severity: After defining the probability, another new column was created named "Severity" which is calculated based on the formula (Pretty, n.d.):

  The Severity of a data breach is the Impact/compromise of a breach multiplied by the Probability (P) that a breach will occur.

  Severity** = **Probability (Likelihood of a breach/ Total occurrence) x Total Record (Impact)

  Due to the large range of values, machine learning algorithms were used. The ML algorithm (Min_max scaler) was used to normalise the data values between 0-1. The severity was labelled as high, medium, low and no severity based on range which is shown below:

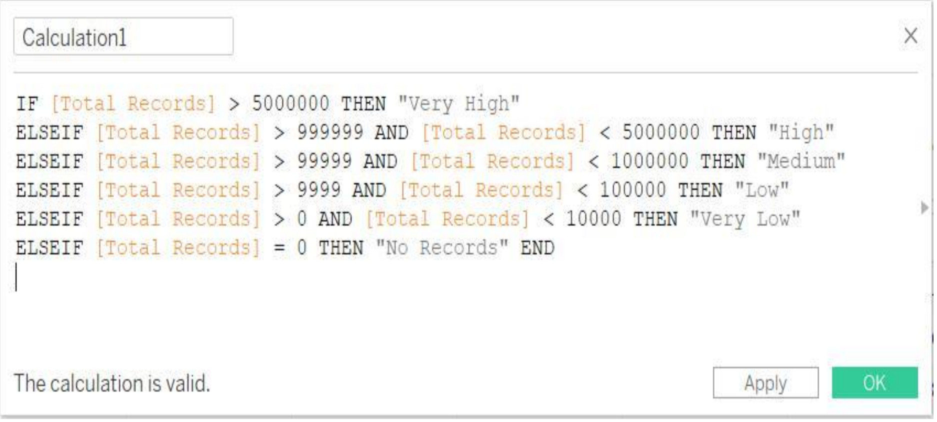


```
In [8]:  min_max_scaler = preprocessing.MinMaxScaler()

In [9]:  x_scaled = min_max_scaler.fit_transform(np.array(df['severity']).reshape(-1,1))
         df_normalized = pd.DataFrame(x_scaled)
         #x_scaled
         df['severity'] = df_normalized

         def sev_lev(x):
             if x == 0:
                 return "No severity"
             elif x>0 and x<0.1:
                 return "low"
             elif x >= 0.1 and x<0.6:
                 return "Medium"
             else:
                 return "High"
```

Figure 3: Python code to add the 'Severity' column

- To calculate the Intensity of the breaches, two new columns were introduced 'Records breached' and 'Intensity'. These were created using Tableau.

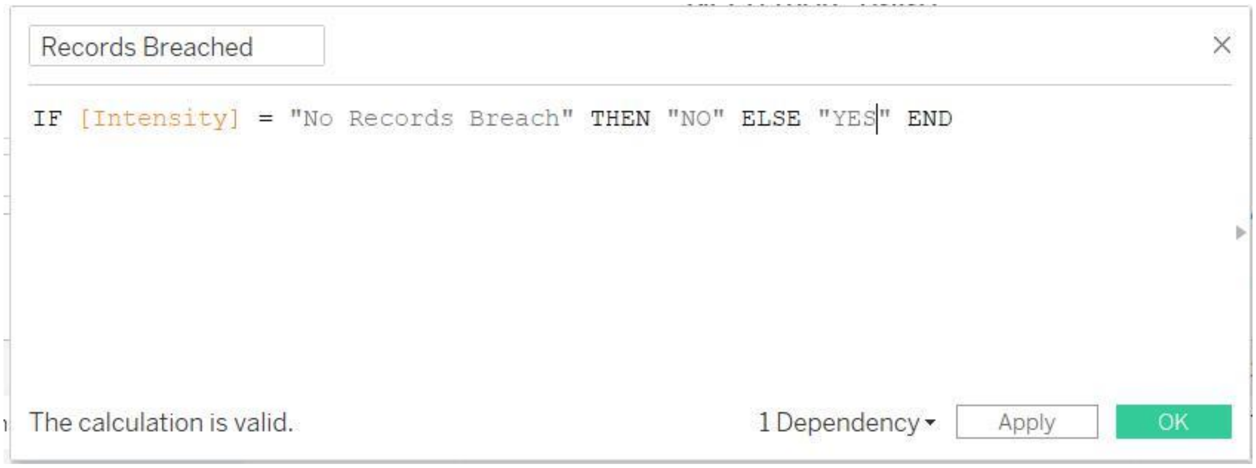
- Intensity: A new column named ‘intensity’ was created using a calculated field. This field divides the breaches records in 6 different segments from very high to very low to no records breached depending on the range records. This calculated field is obtained as follows:



```
IF [Total Records] > 5000000 THEN "Very High"
ELSEIF [Total Records] > 999999 AND [Total Records] < 5000000 THEN "High"
ELSEIF [Total Records] > 99999 AND [Total Records] < 1000000 THEN "Medium"
ELSEIF [Total Records] > 9999 AND [Total Records] < 100000 THEN "Low"
ELSEIF [Total Records] > 0 AND [Total Records] < 10000 THEN "Very Low"
ELSEIF [Total Records] = 0 THEN "No Records" END
```

Figure 4: Tableau code to add the ‘Intensity’ column

- Records breached: This calculated field was created to understand better performing states, since a lot of states were able to save their data from being compromised despite the high count of breaches which makes this column relevant to the performance of these states. It is calculated as follows:

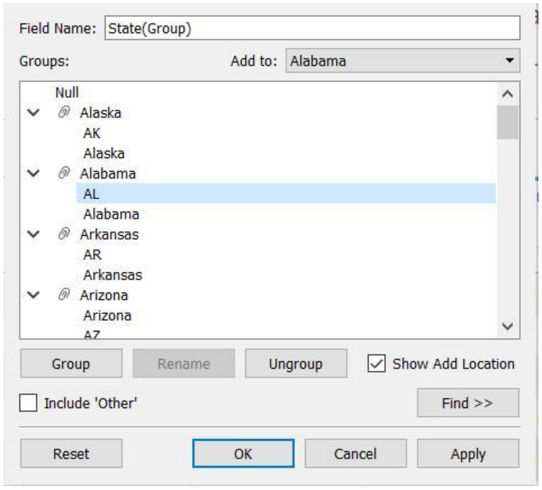


```
IF [Intensity] = "No Records Breach" THEN "NO" ELSE "YES" END
```

Figure 5: Tableau code to add the ‘Records Breached’ column

- State(Group): It considers the states data which will need to be analysed. However, this data required filtering since a lot of states were repeated or mentioned in terms of their short form. This was filtered by using the Roll-up OLAP operation grouping the states together for a better analysis. The state data also contained places that are in the US such as London, Tokyo etc.

All these values as well as the null values were excluded to obtain a refined state column containing only states from the US.



Field Name: State(Group)

Groups: Add to: Alabama

Null
Alaska
AK
Alaska
Alabama
AL
Alabama
Arkansas
AR
Arkansas
Arizona
Arizona
AZ

Group Rename Ungroup ☑ Show Add Location

□ Include 'Other' Find >>

Reset OK Cancel Apply

Figure 6: Tableau Filter to group the 'States' column

Task 1

Visualization 1.1

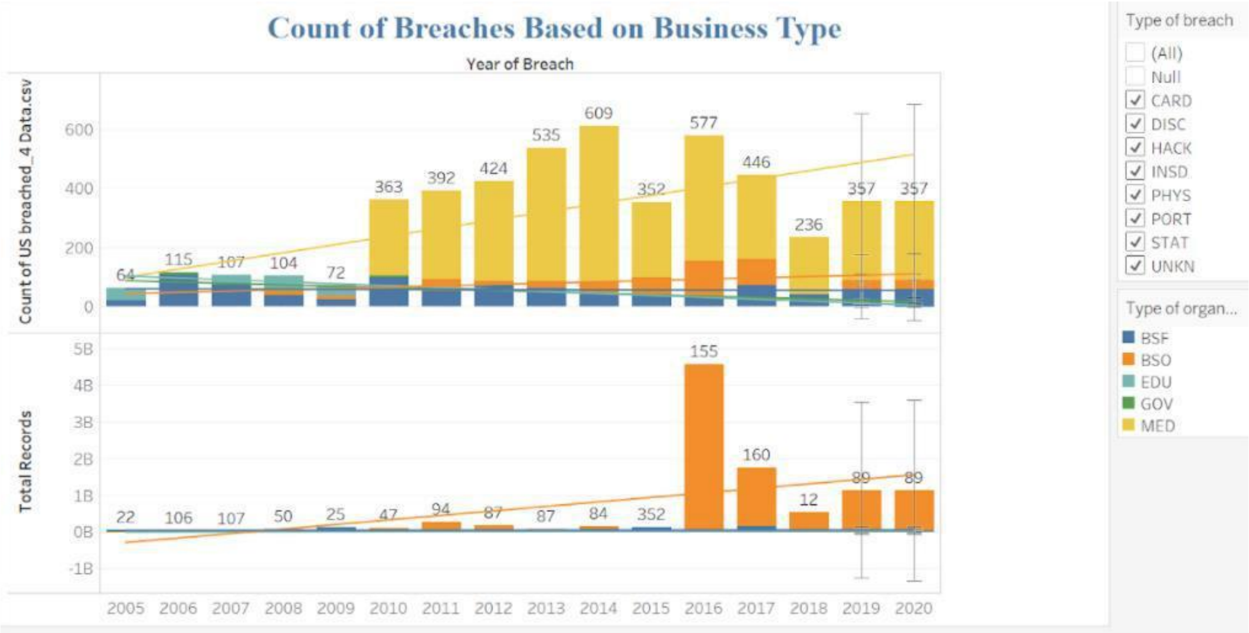


Figure 7: Count of Breaches based on Business type

Implementation

Step 1: Year of the breach was dragged from the dimension in the column section and Total records in the row section, which produced a y-axis of total records, count, and the x-axis of the year of the breach.

Step 2: To make the visualization more clear, filters were added on the type of breach which enabled removing the null values.

Step 3: To get the better insights of the graph, we have dragged the “Types of organisation” into the filter section and shown the filter “Type of Organisation” card.

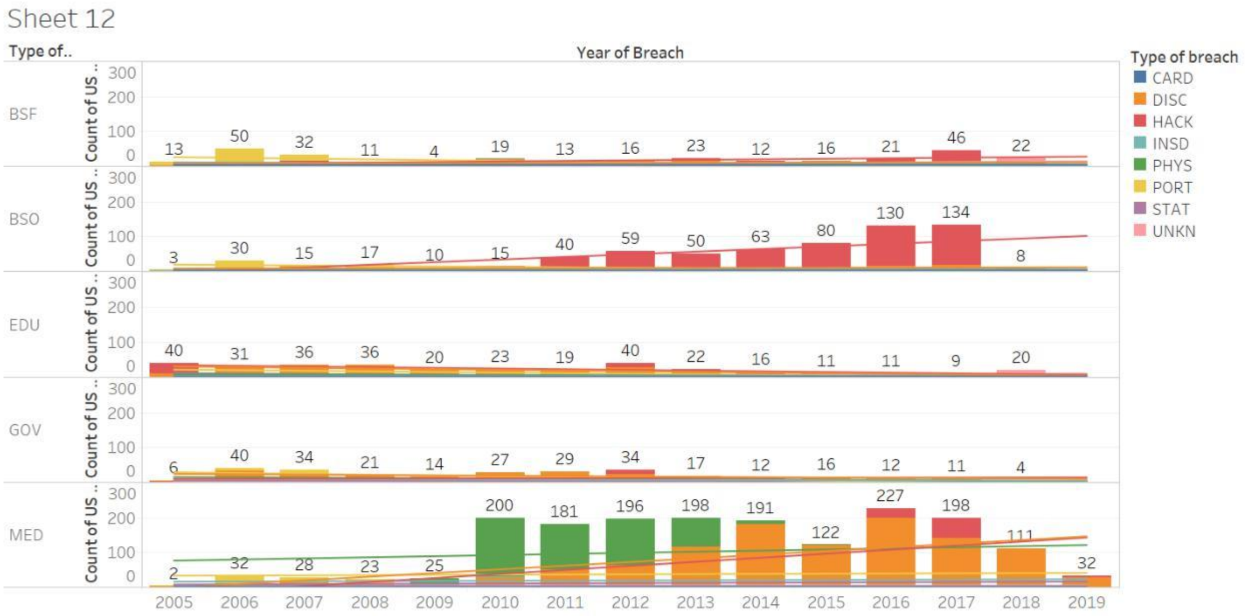
Step 4: The type of breach was dragged into the marks card by putting a colour filter in it which plotted different types of organization in different colours.

Step 5: To show numbers on the graph, Count and total records were dragged from measures to marks section under the label. Show filter was used to enable the checkbox.

Step 6: For displaying the trending lines, visualization was clicked to select the ‘show trend lines’ option for showing different trends for various organizations.

Step 7: Forecast is displayed by clicking on the visualization and selecting the ‘show forecast’ option for showing forecast for various organizations.

Visualization 1.2



Count of US breached_4 Data.csv for each Year of Breach Year broken down by Type of organization. Color shows details about Type of breach. The marks are labeled by count of US breached_4 Data.csv. Details are shown for Type of breach. The data is filtered on Intensity, which excludes Null. The view is filtered on Type of breach and Type of organization. The Type of breach filter excludes Null. The Type of organization filter has multiple members selected.

Figure 8: Intensity over time based on the type of organization

Implementation

Step 1: Year of the breach was dragged from the dimension in the column section and the type of organization in the row section, which produced a y-axis of types of organization and x-axis of the year of the breach.

Step 2: To make the visualization clearer, filters were added on the type of breach & which enabled removing the null values. Moreover, the filter is applied on the intensity that only excludes null values rest all were taken into the consideration.

Step 3: The type of breach was dragged into the marks card by putting a colour filter in it which plotted different types of organization in different colours.

Step 4: The type of breach was dragged into the marks card by putting a colour filter in it which plotted different types of organization in different colours.

Step 5: The intensity was dragged into the filters to remove the null values. After that they were divided into categories: very low, low, medium, high, very high and no records.

Step 6: To show numbers on the graph, Count and total records were dragged from measures to marks section under the label. Show filter was used to enable the checkbox.

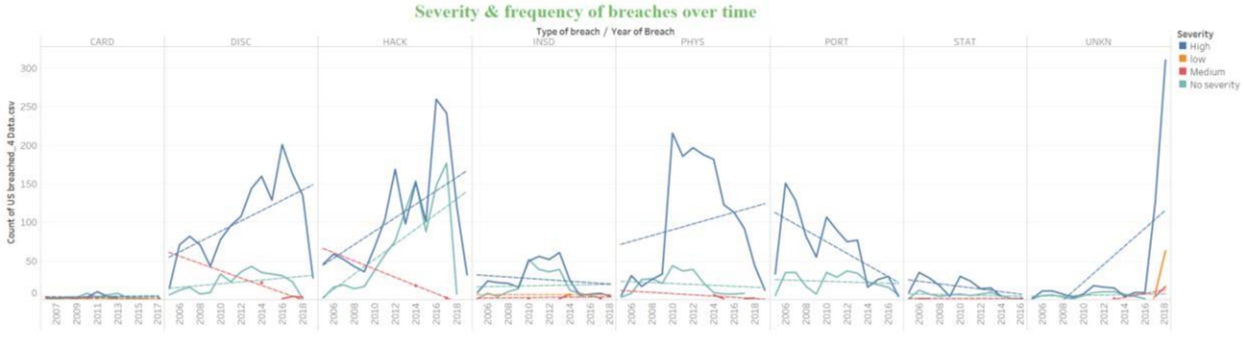
Step 7: For displaying the trending lines, the visualization was clicked to select the 'show trend lines' option for showing different trends for various organizations.

Inference

- By observing two different comparative analyses of the visualization & their Linear trend line (regression Model) we can conclude that intensity & the number of breaches is increasing over the time mainly in MED & BSO organizations.
- From the visualization 1, we can see that the intensity of the count and the total records are increasing at some point but the trend line helps us to clearly infer that the intensity of the breaches are increasing over the time(2005-2018) in MED, BSO organisations and we have also forecasted the intensity of breaches for 2019- 2020.
- Again in order to support our claim we have created another visualisation which clearly shows the increase in the count of intensity over time with respect to each organisation.
- Trend lines are showing the correlation among the explanatory variables and are helping us to forecast the trend. Therefore, on the basis of the trend line we have concluded that we can say that there was an increase in the number of hacks in medical and Business(other) from the year 2005 to 2019.

Task 2

Visualization 2.1



The trend of count of US breached_4 Data.csv for Year of Breach Year broken down by Type of breach. Color shows details about Severity. The view is filtered on Type of breach, which excludes Null.

Figure 9: Severity of breaches over time

Implementation

Step 1: Year of the breach and type of breach was dragged from the dimension in the column section and Total records in the row section, which produced a y-axis of total records and x-axis of the year of the breach.

Step 2: To make the visualization more clear, filters were added on the type of breach which enabled removing the null values.

Step 3: The severity was dragged into the filters by putting a colour filter in which they were divided into categories – high, low, medium, and no severity.

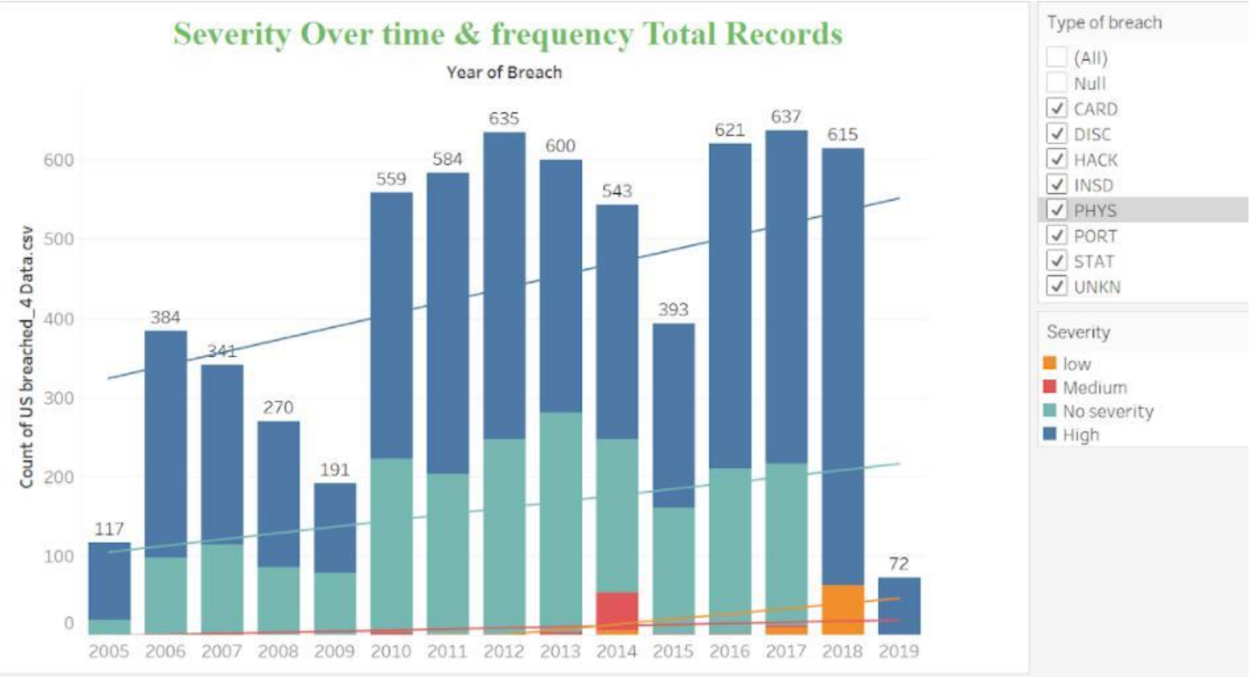
Step 4: To show numbers on the graph, Count and total records were dragged from measures to marks section under the label. We also used a show filter to enable the checkbox.

Step 5: For displaying the trending lines, we clicked on the visualization and selected the 'show trend lines' option for showing different trends for various organizations.

Inference

By looking at the visualization and linear trend line, we can infer that the high severity level has been constantly increasing as the count of breaches increase from 2006-2016, and then there is a sudden fall from 2016-2018 such as HACK, DISC, and PHYS, etc whereas in INSD, PORT & STAT the Linear trend lines shows that the severity & the count of breaches decrease over time. Trending line is the determinant factor because it shows the correlation among the explanatory variables based on R values.

Visualization 2.2



Type of breach

☐ (All)
☐ Null
☑ CARD
☑ DISC
☑ HACK
☑ INSD
☑ PHYS
☑ PORT
☑ STAT
☑ UNKN

Figure 10: Severity over time and frequency of count

Implementation

Step 1: Year of the breach was dragged from the dimension in the column section and Total records in the row section, which produced a y-axis of total records and x-axis of the year of the breach.

Step 2: To make the visualization more clear, filters were added on the type of breach which enabled removing the null values.

Step 3: The severity was dragged into the filters by putting a colour filter in which they were divided into categories – high, low, medium and no severity.

Step 4: To show numbers on the graph, Count was dragged from measures to marks section under the label.

Step 5: For displaying the trending lines, we clicked on the visualization and selected the ‘show trend lines’ option for showing different trends for various organizations.

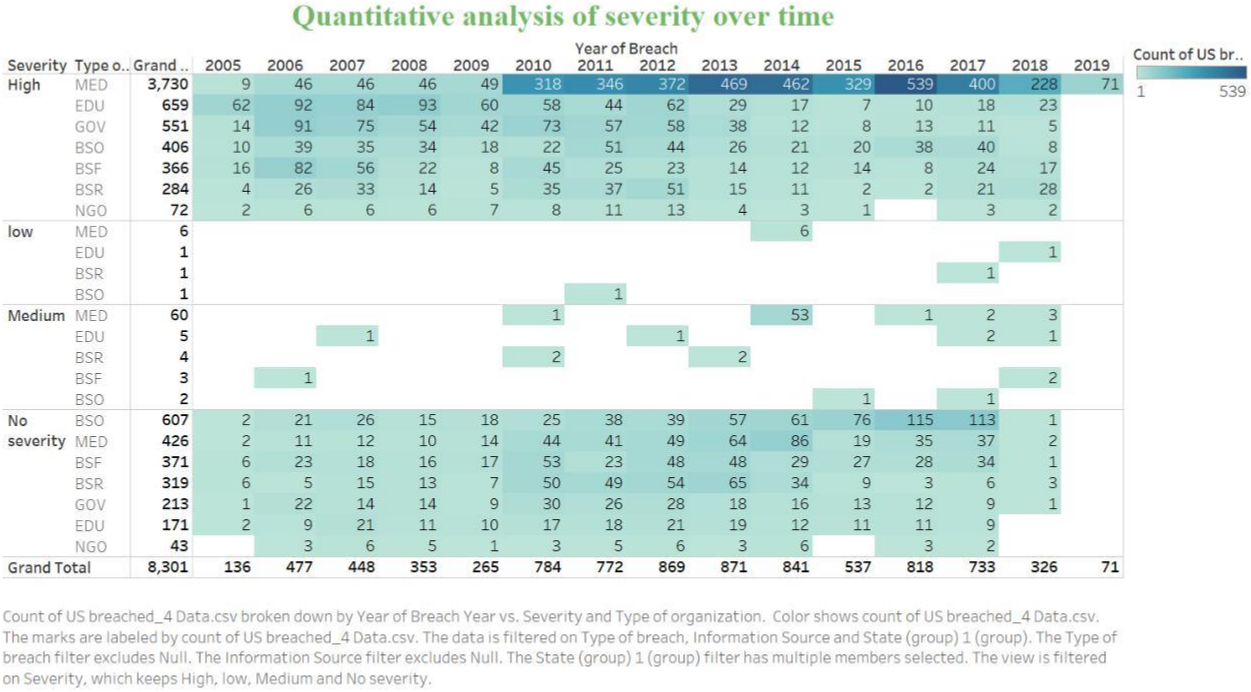
Inference

The graph shows the severity and frequency over the years where every year the severity and frequency is increasing from 2005-2019 with highest for the year 2016 and lowest for 2019. It fluctuates from the year 2005-2009. There is a sudden rise in count of breaches from 2009 to 2010. Overall, the majority of the count of breaches are high severity breaches, whereas the second most breaches occurring are no severity breaches and third most are the medium severity breaches.

Visualization 2.3

To prove our assumption, we considered an example where we are going to see data breaches becoming more frequent and more severe over time in the US. Hence, we are going to see the severity w.r.t organisation over time. For that we have shown two solutions:

Quantitative analysis of data



Quantitative analysis of severity over time

			Year of Breach														
Severity	Type o..	Grand ..	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
High	MED	3,730	9	46	46	46	49	318	346	372	469	462	329	539	400	228	71
	EDU	659	62	92	84	93	60	58	44	62	29	17	7	10	18	23	
	GOV	551	14	91	75	54	42	73	57	58	38	12	8	13	11	5	
	BSO	406	10	39	35	34	18	22	51	44	26	21	20	38	40	8	
	BSF	366	16	82	56	22	8	45	25	23	14	12	14	8	24	17	
	BSR	284	4	26	33	14	5	35	37	51	15	11	2	2	21	28	
	NGO	72	2	6	6	6	7	8	11	13	4	3	1		3	2	
low	MED	6										6					
	EDU	1														1	
	BSR	1													1		
	BSO	1							1								
Medium	MED	60						1				53		1	2	3	
	EDU	5			1					1					2	1	
	BSR	4						2			2						
	BSF	3		1												2	
	BSO	2											1		1		
No severity	BSO	607	2	21	26	15	18	25	38	39	57	61	76	115	113	1	
	MED	426	2	11	12	10	14	44	41	49	64	86	19	35	37	2	
	BSF	371	6	23	18	16	17	53	23	48	48	29	27	28	34	1	
	BSR	319	6	5	15	13	7	50	49	54	65	34	9	3	6	3	
	GOV	213	1	22	14	14	9	30	26	28	18	16	13	12	9	1	
	EDU	171	2	9	21	11	10	17	18	21	19	12	11	11	9		
	NGO	43		3	6	5	1	3	5	6	3	6		3	2		
Grand Total		8,301	136	477	448	353	265	784	772	869	871	841	537	818	733	326	71

Count of US breached_4 Data.csv broken down by Year of Breach Year vs. Severity and Type of organization. Color shows count of US breached_4 Data.csv. The marks are labeled by count of US breached_4 Data.csv. The data is filtered on Type of breach, Information Source and State (group) 1 (group). The Type of breach filter excludes Null. The Information Source filter excludes Null. The State (group) 1 (group) filter has multiple members selected. The view is filtered on Severity, which keeps High, low, Medium and No severity.

Figure 11: Quantitative analysis of severity over time

Implementation

Step 1: Year of the breach was dragged from the dimension in the column section and severity, type of organisation in the row section, which produced a y-axis of severity and x-axis of the year of the breach.

Step 2: To make the visualization more clear, filters were added on the type of breach, information source, states which enabled removing the null values.

Step 3: The count was dragged into the filters by putting a colour filter in which they were divided into categories.

Step 4: To show numbers on the graph, Count was dragged from measures to marks section under the label.

Step 5: Grand total of the count was calculated by selecting 'show column and rows grand total' under the analysis tab.

Inference

- This visualization shows the quantitative analysis of the severity of breaches based on the count of breaches over time. We can infer that the count of breaches in a high level of severity has been increasing over time and the MED organization has maximum counts of 3730 which has been constantly increasing from 2005 to 2017 but there was a sudden fall in count from 2017-2019 whereas the lowest count in high-level severity is 72 of NGO organizations.
- Moreover, this quantitative data is generated based on various factors such as Information of source, states, types of the breach which we have kept under the filter section just to analyse the different patterns over time.

Line graph visualization

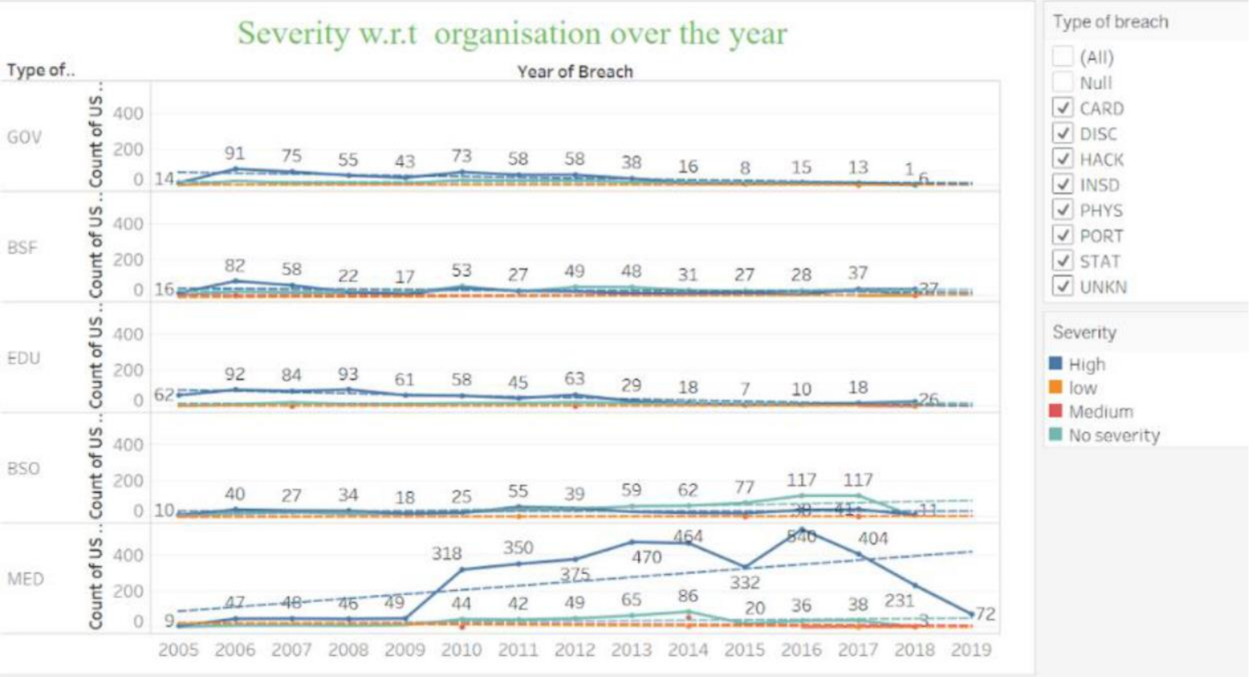


Figure 12: Severity over time through line graph

Implementation

Step 1: Year of the breach was dragged from the dimension in the column section and count, type of organisation in the row section, which produced a y-axis of type of organisation and x-axis of the year of the breach.

Step 2: To make the visualization clearer, filters were added to the type of organisation, type of breach which enabled removing the null values.

Step 3: The severity was dragged into the filters by putting a colour filter in which they were divided into categories – high, low, medium and no severity.

Step 4: To show numbers on the graph, Count was dragged from measures to marks section under the label.

Step 5: For displaying the trending lines, the visualization was selected and 'show trend lines' option was selected for different trends for various organizations.

Inference

To support our quantitative analysis, we took an example to show the severity concerning the organisation over time and for that, we have created a line graph visualization, hence by looking at the graph’s trend line we can infer that the high severity level is increasing over time for example MED organizations.

Conclusion

On the basis of the above inferences we can validate the claim that the severity level has been constantly increasing as the count of breaches increase from 2006-2016 mainly High level, and then there is a sudden fall from 2016-2018

Task 3

The objective of this analysis is to come up with a definition of performance of the states in the US based on the number of breaches and their seriousness. Considering these two factors we have use the following indicators to define the performance:

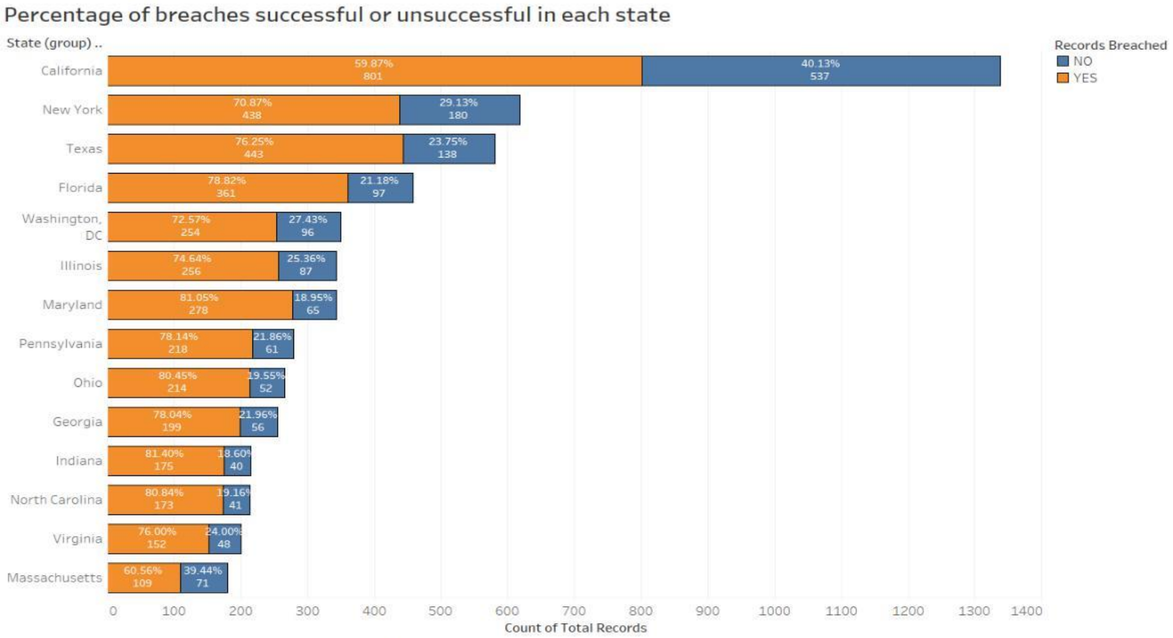
1. Total count or number of breaches: This factor indicates the count of breaches occurring in each and does not consider the seriousness of these breaches. To justify the serious of these breaches we have introduced another indicator called "Intensity"

2. Seriousness: For the seriousness factor of the breaches we have considered 2 factors "severity" and "intensity" to figure out the best performing and worst performing states.

Visualization 3.1

This visualisation plots the states in the US against the count of the number of breaches.

This visualisation also considers another factor that shows how many of those breaches were successful in compromising the data.



Count of Total Records for each State (group) 1 (group). Color shows details about Records Breached. The marks are labeled by % of Total Count of Total Records and count of Total Records. The data is filtered on % of Total Count of Total Records, which includes everything. The view is filtered on Records Breached, State (group) 1 (group), Exclusions (Records Breached,State (group) 1 (group)) and Inclusions (Records Breached,State (group) 1 (group)). The Records Breached filter keeps NO and YES. The State (group) 1 (group) filter excludes Arizona. The Exclusions (Records Breached,State (group) 1 (group)) filter keeps 119 members. The Inclusions (Records Breached,State (group) 1 (group)) filter keeps 28 members.

Figure 13: Percentage of breaches successful or unsuccessful in each state

Implementation

Step 1: Total Records and State(Group) were dragged and dropped in the column and row which resulted in forming the x-axis and y-axis.

Step 2: Only states with total records more than or equal to 200 were kept and the other states were excluded with a filter on State(Group)

Step 3: No records breached was dragged and dropped in the colour section in the Mark card such that:

Blue - No records were compromised.

Orange - The records were compromised by these breaches.

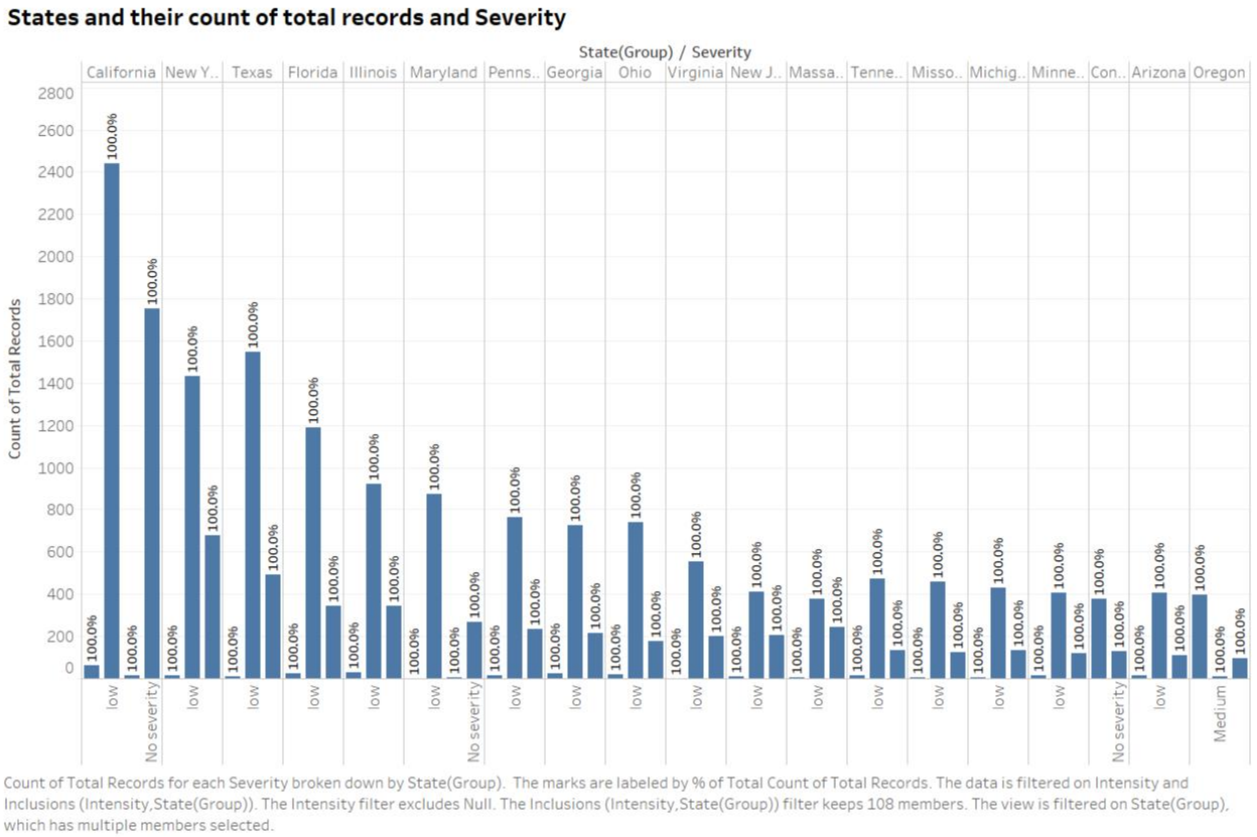
Step 4: The data was filtered on percentage of Total Count of Total Records.

Inference

Although the first 5 states in the visualization had the highest number records for breaches, they were successful in protecting their records from getting breached as well. It also helped in shortlisting the states that are relevant for the analysis, since they represent the maximum amount of information available in the dataset.

Visualization 3.2

This visualisation takes the analysis a step further, by analysing the seriousness of the breaches based on the severity of the breaches which is defined in the previous Tasks 1 and 2.



States and their count of total records and Severity

Count of Total Records for each Severity broken down by State(Group). The marks are labeled by % of Total Count of Total Records. The data is filtered on Intensity and Inclusions (Intensity,State(Group)). The Intensity filter excludes Null. The Inclusions (Intensity,State(Group)) filter keeps 108 members. The view is filtered on State(Group), which has multiple members selected.

Figure 14: State and Severity of breaches

The different column represents the different levels of severity of breaches from high to very low.

Implementation

Step 1: State (Group) and Count of Total Records were dragged and dropped in the column and row which resulted in the formation of x-axis and y-axis.

Step 2: Only states with total records more than or equal to 200 were kept and the other states were excluded with the exclude option on State(Group).

Step 3: The marks are labelled by percentage of Total Count of Total Records.

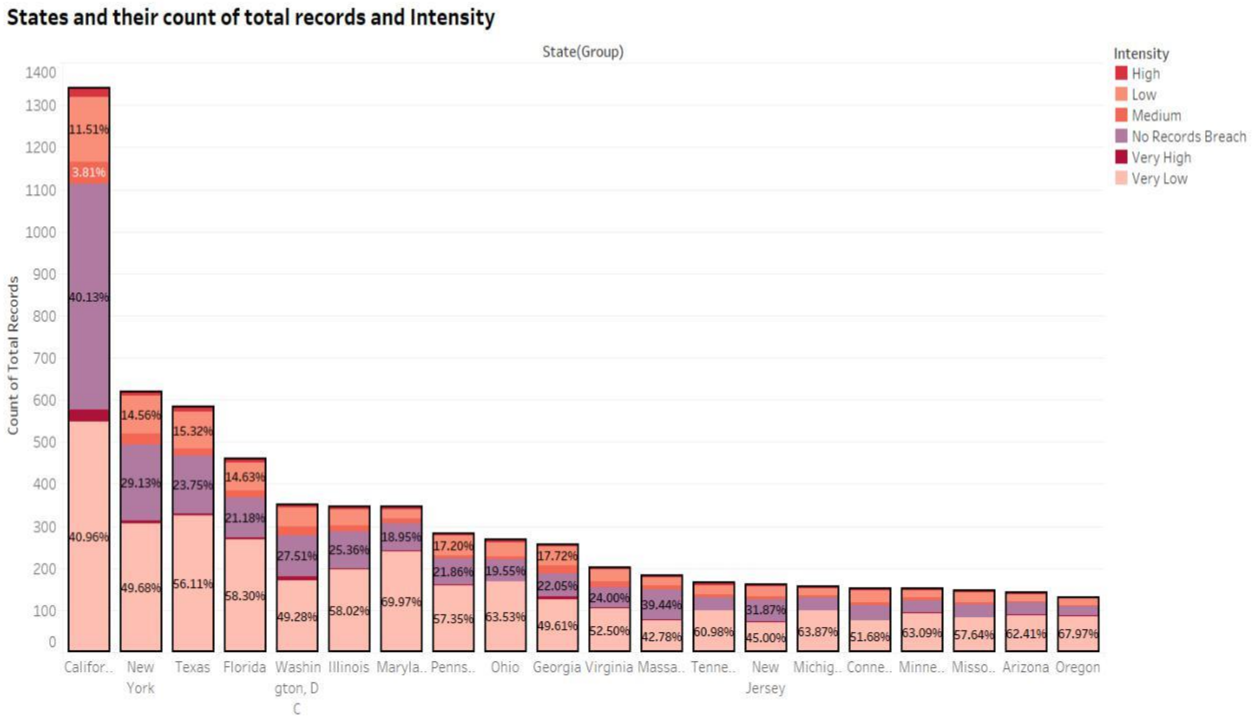
Step 4: Count of Total Records for each Severity was broken down by State(Group)

Inference

From the above visualisation we can easily infer that severity does not play a vital role in conducting further analysis as most of the breaches occurring in these states are very low in severity or have no severity. Therefore, for further analysis of the seriousness of breaches we can ignore the severity of the breaches and focus solely on the intensity of these breaches.

Visualization 3.3

This visualization analyses the seriousness of breaches based on the intensity of the breaches. The intensity of these breaches is based on the evaluation of the table “Total records breached”, this factor has been considered to define the seriousness of the breaches as it justifies the amount of data that these breaches were capable of comprising.



Count of Total Records for each State(Group). Color shows details about Intensity. The marks are labeled by % of Total Count of Total Records. The view is filtered on State(Group), Intensity and Inclusions (Intensity,State(Group)). The State(Group) filter keeps 49 of 65 members. The Intensity filter excludes Null. The Inclusions (Intensity,State(Group)) filter keeps 114 members.

Figure 15: States and their count of total records and intensity

Implementation

Step 1: State(Group) and Count of Total Records were dragged and dropped in the column and row which resulted in the formation of x-axis and y-axis.

Step 2: Intensity was dragged and dropped on a filter to differentiate between intensities.

Step 3: Count was dragged and dropped in the Label of the Marks card which showed the percentage in the visualization.

Step 4: State(Group) was filtered out for selecting the relevant states used in the previous visualisation using the keep only option.

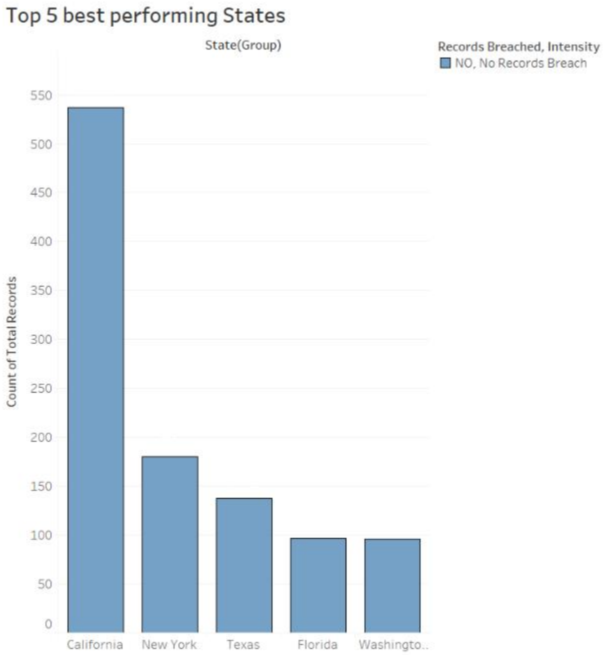
Inference

The inferences obtained from this visualisation helps to identify the top 5 best and worst performing states. For example, although the states of California, New York or Texas have high occurrences of the breaches, the number records breached are very low or have very low intensity. On the other hand, the states with count of records breached such as Maryland, Georgia, etc., were not able to protect their data from being accessed. Therefore, this visualisation can be used to analyse the seriousness of the breaches based on these inferences and the results from the first visualisation.

Visualization 3.4

The visualisation will display the best performing states based on the following definition:

“The states that we have considered best performing are the states with any number of counts of breaches with either no data being compromised by those breaches or a low intensity of the breach meaning despite its occurrence the breach was not able to access a lot of records.”



Count of Total Records for each State(Group). Color shows details about Records Breached and Intensity. The marks are labeled by count of Total Records and % of Total Count of Total Records. The data is filtered on % of Total Count of Total Records, which includes everything. The view is filtered on Exclusions (Records Breached,State (group) 1 (group)), Records Breached, State(Group), Inclusions (Records Breached,State (group) 1 (group)) and Intensity. The Exclusions (Records Breached,State (group) 1 (group)) filter keeps 48 members. The Records Breached filter keeps NO. The State(Group) filter excludes Arizona. The Inclusions (Records Breached,State (group) 1 (group)) filter keeps 5 members. The Intensity filter keeps 7 of 7 members.

Figure 16: Top 5 best performing states

Implementation

Step 1: State(Group) and Count of Total Records were dragged and dropped in the column and row which resulted in the formation of x-axis and y-axis.

Step 2: Only the top 5 states from the State(Group) were filtered out.

Step 3: Records breached, and intensity were dragged and dropped in the filter section.

Step 4: The records breached data used for this visualisation only contains values with no records breached.

Inference

In conclusion we can say that the following are the best performing states:

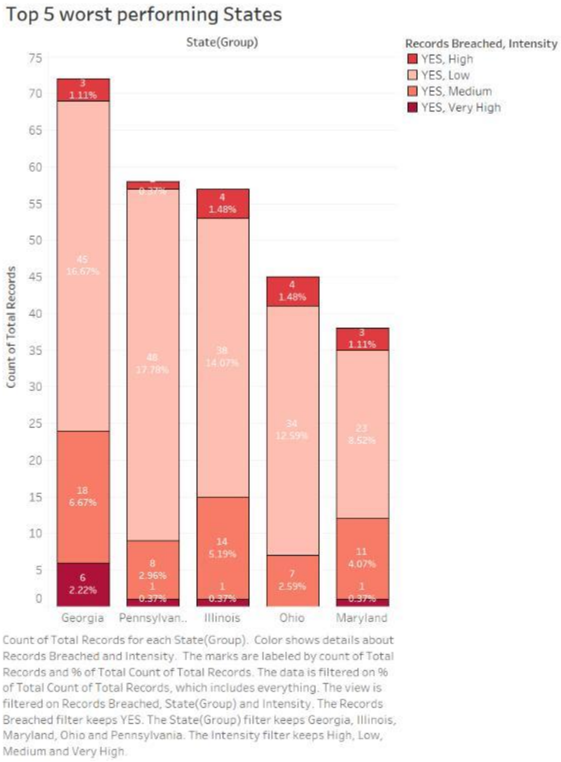
1. California
2. New York

3. Texas
4. Florida
5. Washington

Since these states were successfully able to perform well by protecting their data from the breaches meaning the intensity of the breaches in these states was very low despite the high volume of breaches occurring in these states.

Visualization 3.5

This visualization displays the worst performing states based on the following definition: “The states which have identified as worst performing are the ones with any count of number of breaches having high intensity of breaches occurring on their records and having a lot of their data being compromised, meaning these states were not able to protect their data from being exposed to these breaches.”



Count of Total Records for each State(Group). Color shows details about Records Breached and Intensity. The marks are labeled by count of Total Records and % of Total Count of Total Records. The data is filtered on % of Total Count of Total Records, which includes everything. The view is filtered on Records Breached, State(Group) and Intensity. The Records Breached filter keeps YES. The State(Group) filter keeps Georgia, Illinois, Maryland, Ohio and Pennsylvania. The Intensity filter keeps High, Low, Medium and Very High.

Figure 17: Top 5 Worst Performing states

Implementation

Step 1: State(Group) and Count of Total Records were dragged and dropped in the column and row which resulted in the formation of x-axis and y-axis.

Step 2: Only the top 5 states from State(Group) were displayed using a filter on that field.

Step 3: Records breached were dragged and dropped in the filter which in this visualisation only contains values with records breached and excludes the data where no records were breached.

Step 4: Intensity was dragged and dropped in the Colour of the Marks card which results in obtaining values with intensity ranging from high to low levels from its 6 levels.

Inference

In conclusion we can say that the following are the worst performing states:

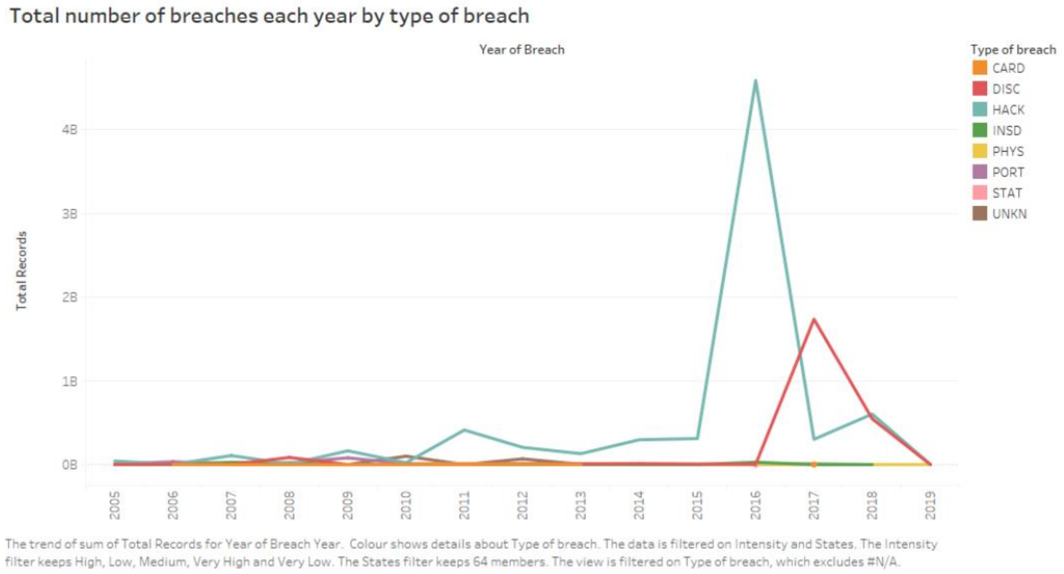
1. Georgia
2. Pennsylvania
3. Illinois
4. Ohio
5. Maryland

Since they were unable to perform well in terms of protecting their data from the breaches meaning the intensity of the breaches in these states was very high despite the low volume of breaches occurring in these states.

Task 4

Visualization 4.1

The visualization finds the trend in the types of breaches in the United States of America from 2005 to 2019.


Total number of breaches each year by type of breach

The trend of sum of Total Records for Year of Breach Year. Colour shows details about Type of breach. The data is filtered on Intensity and States. The Intensity filter keeps High, Low, Medium, Very High and Very Low. The States filter keeps 64 members. The view is filtered on Type of breach, which excludes #N/A.

Figure 18: Total number of breaches each year by type of breach

Implementation

Step 1: Year of breach was dragged in the column and Total records in the row, which produced a y-axis of total records and x-axis of year of breach. To understand the trend in the breaches line graph was used.

Step 2: Type of breach was dragged into the Marks card by putting a color filter in it which plotted different types of breaches in different colors.

Step 3: To make the visualization clearer, filters were added on the type of breach which enabled in removing the null values.

Inference

HACK - hiked in 2015 lowered in 2017, the highest record of breaches throughout 2005-2019 was the HACK attack. The total number of records was more than 4 billion.

DISC - Saw a hike in 2017 which was around 2B and the number of breaches gradually started declining in 2017.

CARD- fraud involving debit and credit cards were noticed only from 2006-2013. There was only one record of CARD breach in 2017.

INSD- two peaks were noticed, which were in 2007 and 2016. The maximum records breached were during the second hike which was almost 30M.

PHYS- It was constant from 2005-2019 with two minor peaks in 2011 and 2014.

PORT- The number of records increased in 2009 and declined dramatically afterwards till 2017

STAT- from 2005- 2014 it has been almost steady

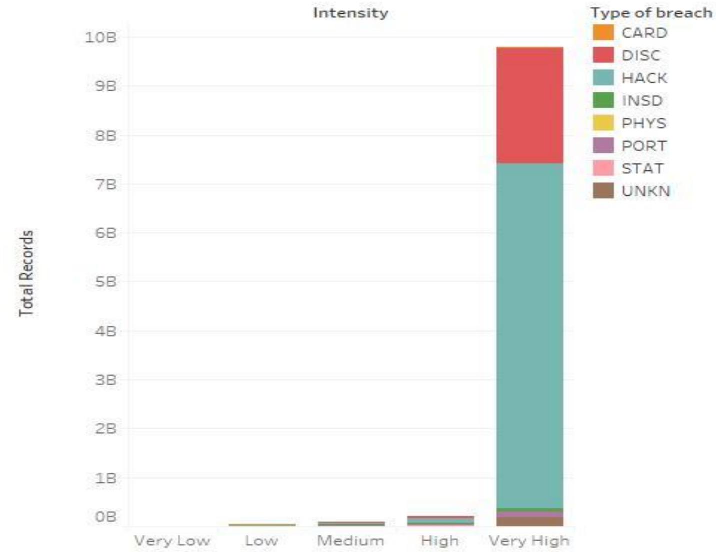
UNKN- There were two major peaks, one in 2010 and the other in 2012. It eventually declined dramatically after 2012.

Hence, the highest types of breaches in the USA are hacking by an outside party or infected by malware.

Visualization 4.2

This visualization finds the intensity of the types of breaches

Intensity of the Breaches by type of Breach



Sum of Total Records for each Intensity. Colour shows details about Type of breach. The data is filtered on States, which keeps 64 members. The view is filtered on sum of Total Records, Type of breach and Intensity. The sum of Total Records filter keeps non-Null values only. The Type of breach filter excludes #N/A. The Intensity filter keeps High, Low, Medium, Very High and Very Low.

Figure 19: Intensity of the Breaches by the type of Breach

Implementation

Step 1: Intensity was dragged and dropped from the dataset to the column section which gives the intensity x-axis.

Step 2: The Sum of the total number of records was set at the row section which gives Total Records y-axis.

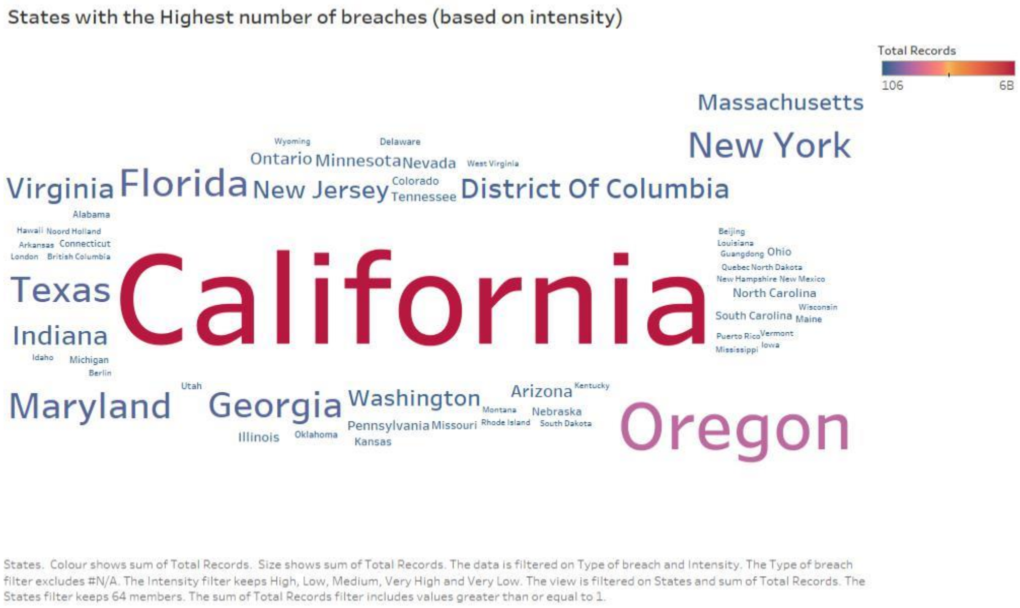
Step 3: The types of breach are highlighted with different colors according to the breach type to create a distinction.

Inference

The highest intensity among all the types of breaches is the HACK breach which has a record of 7 Billion. Amongst all the various intensities of the breaches which took place over the years, the intensity which affected the most is very high intensity. The intensity type was calculated by keeping in mind the number of records lost during the breach.

Visualization 4.3

This visualization finds out the breach at a geographic region with respect to the intensity of breach.



States with the Highest number of breaches (based on intensity)

States. Colour shows sum of Total Records. Size shows sum of Total Records. The data is filtered on Type of breach and Intensity. The Type of breach filter excludes #N/A. The Intensity filter keeps High, Low, Medium, Very High and Very Low. The view is filtered on States and sum of Total Records. The States filter keeps 64 members. The sum of Total Records filter includes values greater than or equal to 1.

Figure 20: States with the highest number of breaches based on intensity

Implementation

Step 1: The States were dragged and dropped in the text filter in the Marks Card.

Step 2: The Sum of Total Records was dragged and dropped in the color and size filter in the Marks Card

Step 3: To highlight the level of intensity a diverging color set was chosen which presents the state with the highest intensity with a bright red.

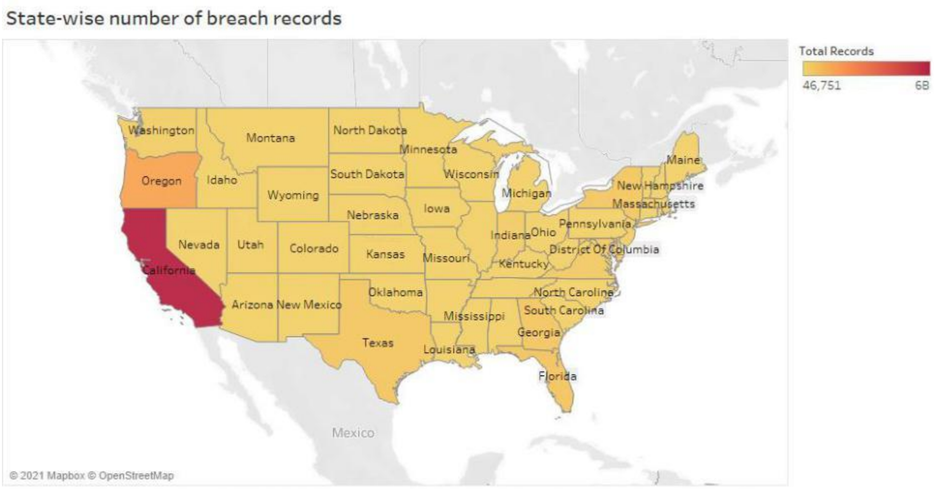
Step 4: To group the states according to the intensity, Intensity was dragged and dropped in the Filters tab to include only very high, high, low, very low and medium intensity.

Inference

This visualization was constructed to give us the name of the state where most of the breaches took place. California has the highest number of breaches which can be clearly seen from the visualization.

Visualization 4.4

This visualization finds the state with the highest number of breach records



Map based on Longitude (generated) and Latitude (generated). Colour shows sum of Total Records. The marks are labelled by States. Details are shown for States. The data is filtered on Type of breach and Intensity. The Type of breach filter keeps DISC, HACK, PHYS, PORT and UNKN. The Intensity filter keeps High, Low, Medium, Very High and Very Low. The view is filtered on States, Latitude (generated) and Longitude (generated). The States filter keeps 64 members. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Figure 21: State- wise number of breach records

Implementation

Step 1: Generated Longitude and the generated latitude was dragged and dropped in the column and row respectively.

Step 2: Filter was added to both Longitude and Latitude to remove the null values by going in the Specials tab.

Step 3: States were dragged and dropped in the Marks Detail Card. Total Records was dragged and dropped in the Marks Color Card. This will finally generate the map of the world with the states with recorded breaches in different colors depending on the number of records.

Step 4: Filter was added on States to remove null values and states that are not a part of the United States.

Inference

The visualization shows that California is the state with the highest number of breach records as it is the state that stands out the most. The color in the map fades with the states with less recorded breaches. The visualization also clearly shows that Oregon is the second state with the highest number of breaches recorded.

Dashboard

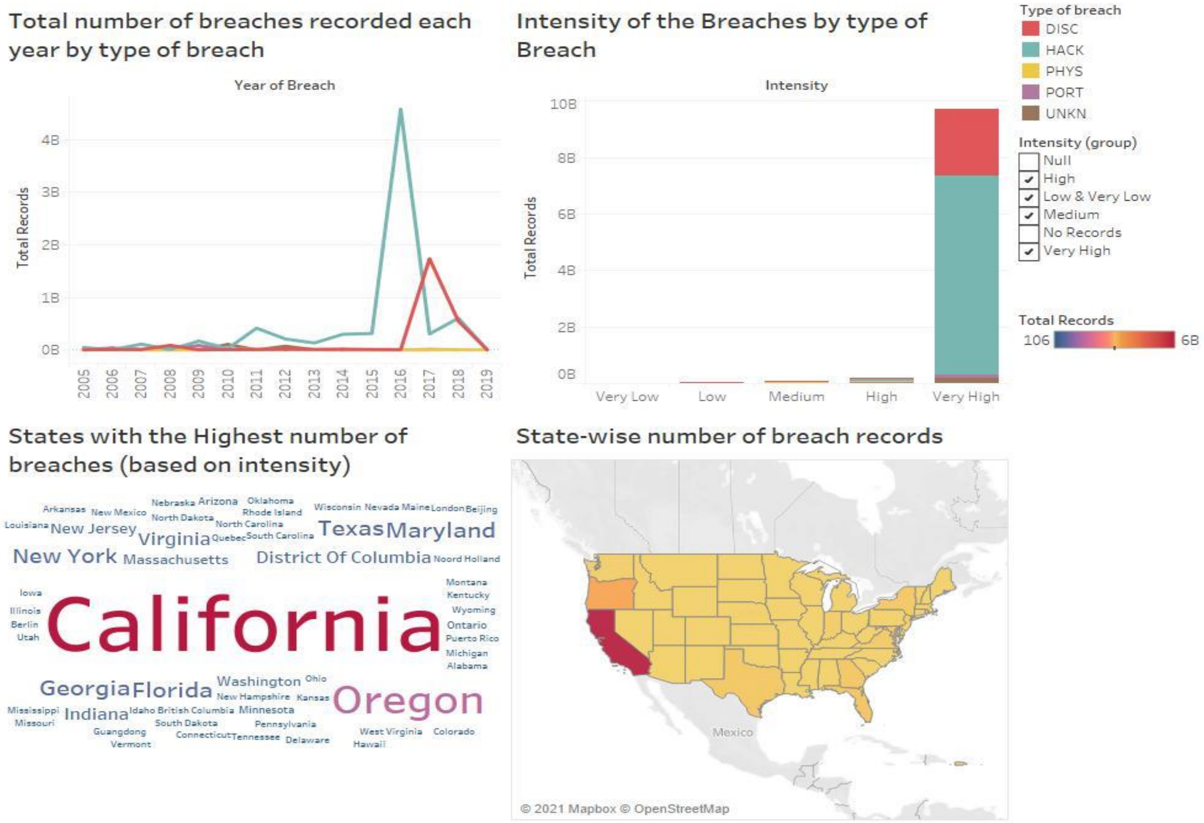


Figure 22: Dashboard for US Breaches Dataset

From the dashboard it can be seen that the breaches have been increasing from 2015 to 2018 and HACK breaches have the highest intensity. The state which has recorded the highest number of breaches is California.

References

Data Breaches Tableau Visualization. (n.d.). From Kaggle: https://www.kaggle.com/xvivancos/data-breaches-tableau-visualization

Pretty, T. (n.d.). *Calculate risk in under 2 mins*. From Cipherpoint: https://cipherpoint.com/blog/calculaterisk/

Sundareswaran, V. (2018). STUDY OF CYBERSECURITY IN DATA BREACHING.