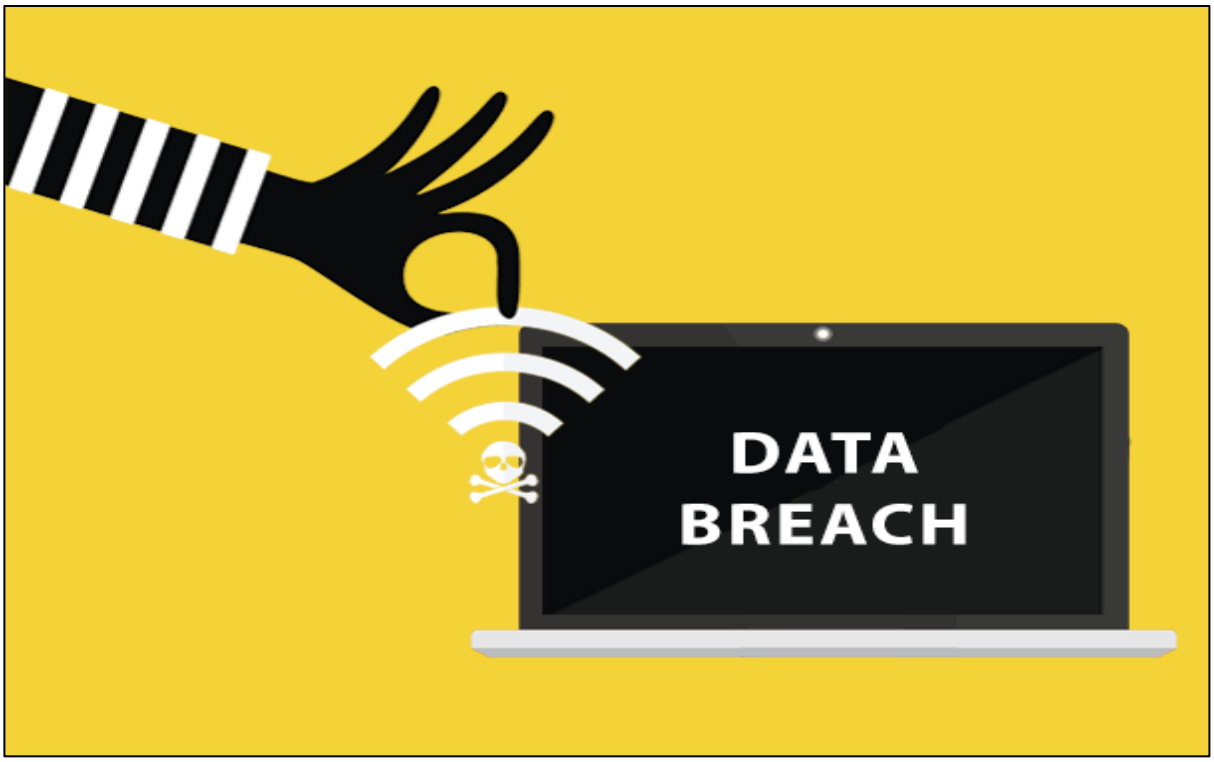
FIT5147 DATA EXPLORATION AND VISUALISATION
DATA EXPLORATION REPORT

"DATA BREACHES IN ORGANISATIONS AND SECTORS ANALYSIS"



Name: **Arushi Tejpal**
Student ID: **28130006**
Tutorial: **Friday 4pm**

TABLE OF CONTENTS:

1. Introduction

Data breaches is the intentional or unintentional release of secure or private data information to an untrusted environment. Data breaches can occur through hacks, poor security, lost device or an inside job.

This theft or loss of digital media has become a major concern for individual privacy and the trust within consumers may soon be lost. The modern world is growing with information and we are all constantly exposing our personal information to the internet, its important to know where these data breaches can come from and how they occur.

I have found this dataset from informationisbeautiful.net;
It contains information about companies names and the records lost from each company. I have used Tableau, R and Excel to analyse the dataset.

We will address **five main questions** which will help understand where the breaches occur, when they occur and why the occur.

1. **What were the most common causes for data breaches to occur?**
2. **What sector were the most data breaches found?**
3. **What Year from 2004 - 2020 did the most data breaches occur?**
4. **How sensitive the data is that's being lost?**
5. **What country has the most data breaches?**

2. Data Wrangling

To answer these questions we need to fist better understand our data and find ways to clean our data to suit the questions we answer.
I will be making 10 different data visualisation graphs to answer the questions.
I used Tabular data and spatial data to analyse my questions.

Tools Used:

1. **R studio**
2. **Tableau Public**
3. **Excel**

DATA CLEANING and TRANSOFRMATION

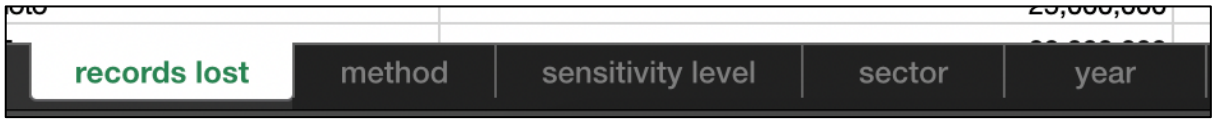
1. **Cleaning unwanted Columns.**

   First I deleted unnecessary such as column X , because it was an empty column in our data set. This was done through excel.

2. Extraction of new columns - Transformation

From my original data I extracted specific columns and created 5 separate sheets through excel, the sheets were based on: records lost, method, sensitivity level, sector and year. This is done so its easy to extract graphs on to Tableau. As seen below:

Fig. 1:

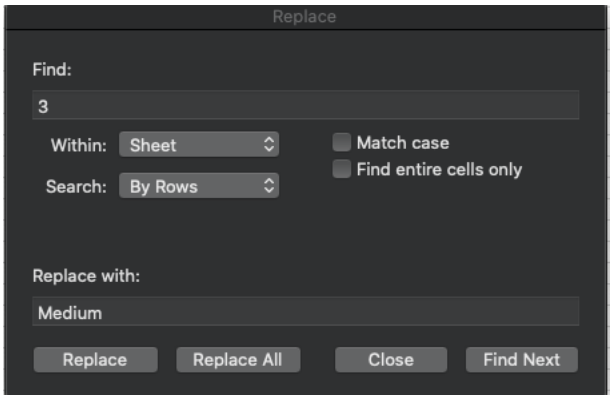


3. Replacing data in column "Sensitivity level"

The "data sensitivity" column was originally numbered from 1 to 5.

I had changed this to make a range of very low to very high as it will make a clear representation about data sensitivity. This had been done through excel by filtering each number and replacing it with the correct sensitivity level. Seen in fig.

Fig. 2:



4. Creating a new Data frame

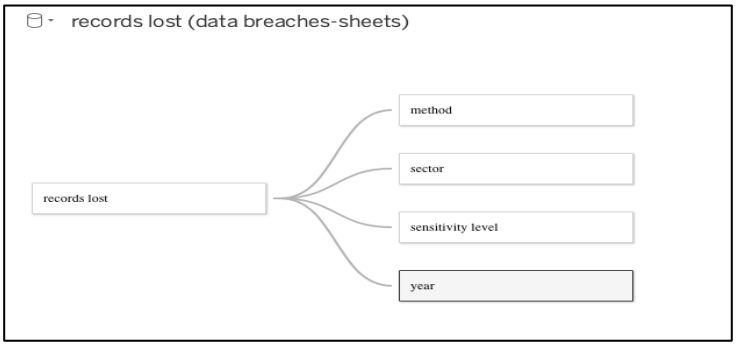
For getting a data regarding what country had the most breaches I had to manually collect data about where the entity was located. I first made the companies (entity column) unique through using R. This made it easier to go through each company and find their location.

Fig 3:

```
entity_name<-unique(r$entity)
```

5. Joining data frames

To join the data frame I had a many to one relationship with the records lost column and method, sensitivity level, sector and year. This was done using Tableau.



3. Data Checking

Error 1: Null values:

Removing null values

Removing NaN values was done through the use of R. there were many companies which did not have any records lost listed and did not have a location belonging to them. These columns had to be removed to create accurate results. As seen below: "OVH" does not have any records listed this row is useless as no meaningful conclusion can be made.

Fg. 5.

Crescent Health Inc., Walgreens		100,000
Florida Department of Juvenile Justice		100,000
Advocate Medical Group		4,000,000
OVH	French Internet host	
Apple		275,000

Error 2: Un-reliable values:

Removing multiple entity columns

I also had to remove rows which had multiple entities written such as "LinkedIn, eHarmony , Last.f" This was done because all the entities have to be in separate rows. As seen in below: fig 6.

Last.fm	43,500,000
LinkedIn, eHarmony, Last.fm	8,000,000
Gaming	3,000,000

Error 3: Duplicate values

Checking duplicate values and their removal.

I further removed duplicate "Entity" column, and sorted the data according to the year.

Error 4: Removing Values

Removing extra sectors

Some of the entities had two sectors attached to them. When I make a bar graph this makes extra rows in tableau.

Fig. 7.

Yahoo	web
SnapChat	web, tech
University of Delaware	academic
Central Hudson Gas & Electric	energy

Error 5: Removing Null values

Removing Null value for countries.

Some of the entities did not have country location

This was removed using R. fig 7.1:

US Marshals Service	United States of
db8151dd	
EasyJet	United States of
Microsoft	United States of
Dutch Government	Netherlands
Virgin Media	United Kingdom
Boots Advantage Card	United Kingdom
Tesco Clubcard	United Kingdom
Marriott Hotels	United States of
Zoom	United States of
Israeli government	Israel
MGM Hotels	United States of
Buchbinder Car Rentals	Germany
Wawa	United States of
Desjardins Group	Canada
US Customs and Border Prote	United States of
Quest Diagnostics	United States of
Australian National University	Australia
Canva	Australia
First American Financial Corp	United States of
Chtrbox	India
WiFi Finder	

4. Data Exploration

1) What were the most common causes for data breaches to occur?

Answer:

- From figure 8. We canobserve that the number of records lost through lost device and inside job is very low compared to losing records through poor security and being hacked.
- The most common causes of data breaches was by poor security with over 10 billion records.
- 1.65% and 1.06 % records are lost thorugh inside job and lost device (respectively)
- 62% of the records are lost through poor security as seen through figure 9.
- We can also observe that the sector which has lost most of its data through poor security is the tech sector. And the sector which has been hacked the most is the web sector. The majority of accidental "oops!" leaks that occur is through the web industry.

I chose the bar graph and the pie graph as a form of visulaisation for this question . The pie graph shows the percentage of records lost per sector, this is a good representation of the overall view. The bar graphs gives a more detailed analysis as the different lengths represent the number of records lost.

Figure 8: (on the left) Bar graph of Relationship between records lost and method, divided by each sector.

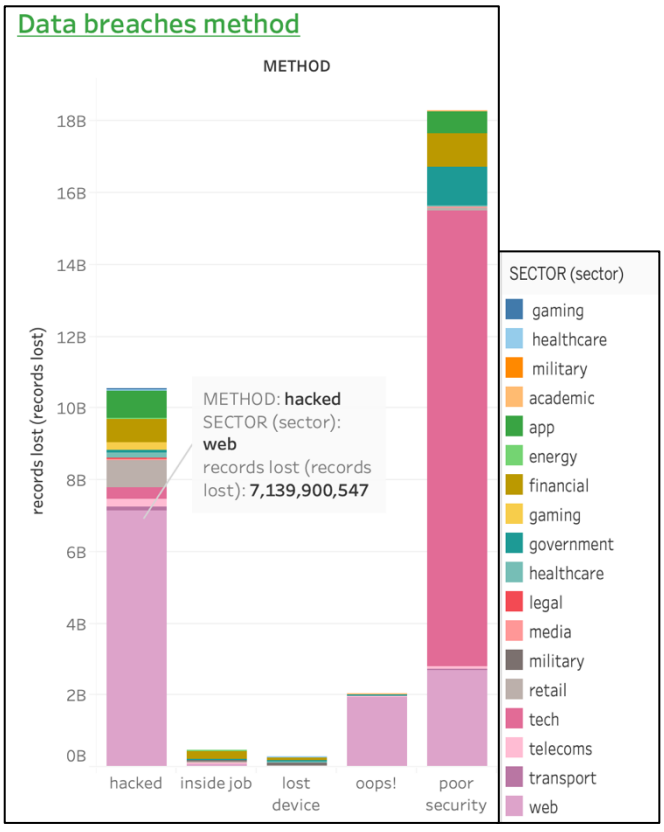
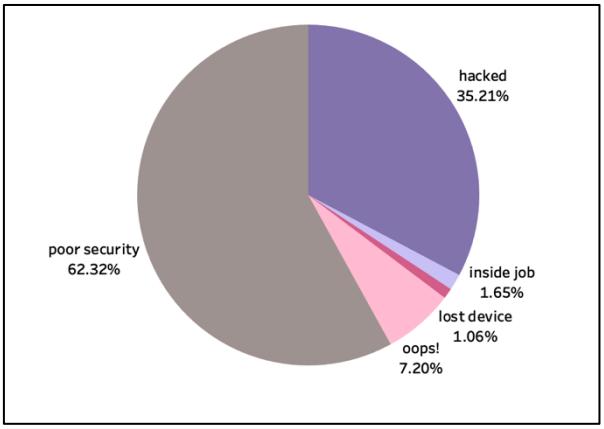


Figure 9. (on the right) Pie Graph. Percentage of records lost by each method.



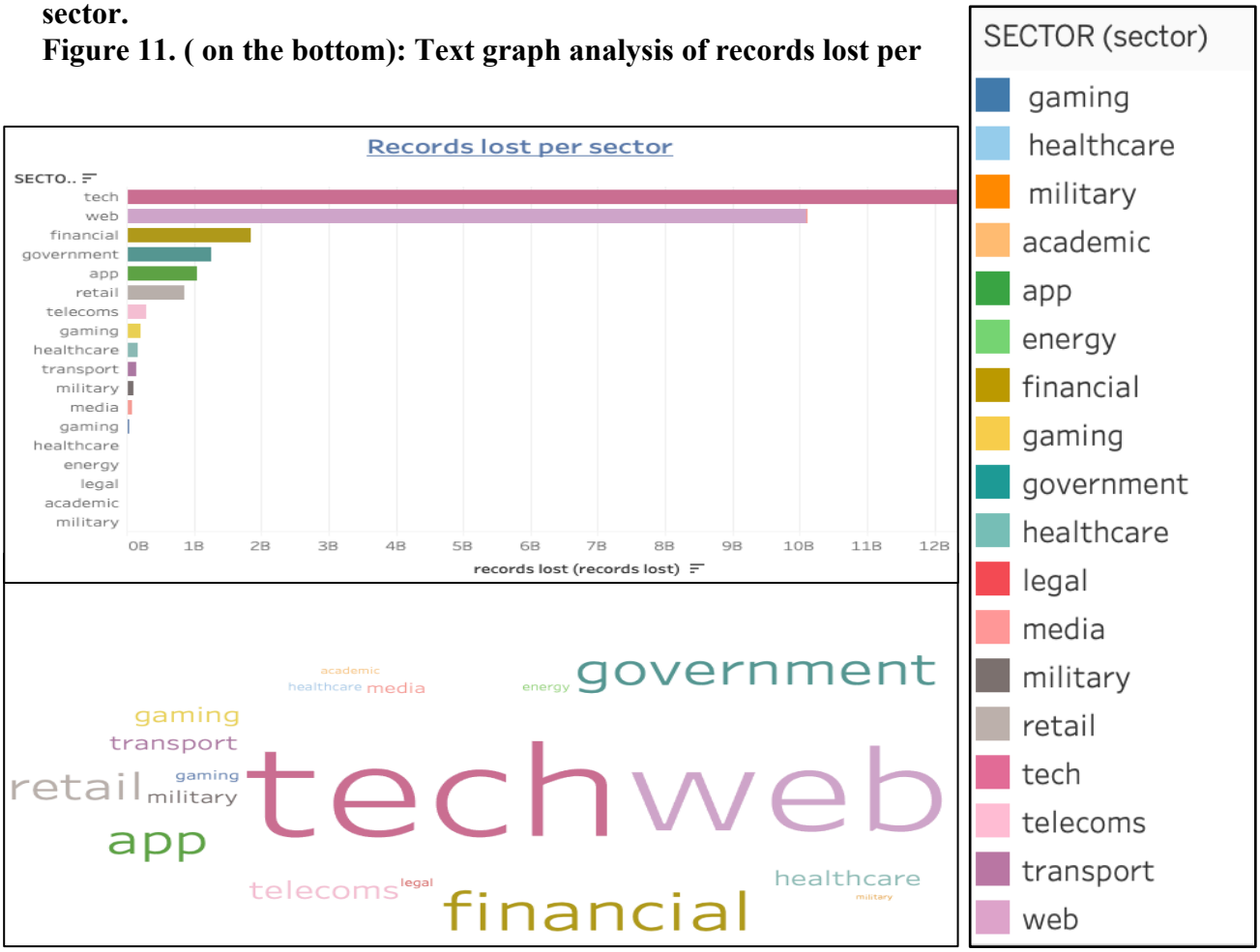
2) What sector were the most data breaches found?

Answer:

- By using Tableau I was able to create two visualisation analysis of the sector vs records lost relationships. There were a few null values which I had been deleted previously by R studio. Some of the companies/ entities had two sectors written next to them which caused an extra bar, this was also removed and filtered out using excel.
- From figure. 10 and 11. we can conclude that the most data breaches occurred in the tech industry and the web sector not too far along. The least breaches occurred in the military sector which makes sense as their confidential data would be hard to breach. The tech and web sector are responsible for the most data leaks maybe due to their large online presence as more people are signing up, thus, more leaks and breaches would occur.

I chose the word cloud analysis as it gives a visual representation of the frequency of records lost in each sector. The bar graph gives a more detailed analysis showing the number of records lost.

Figure 10. (on the top) : Bar graph of records lost per sector sector.
Figure 11. (on the bottom): Text graph analysis of records lost per

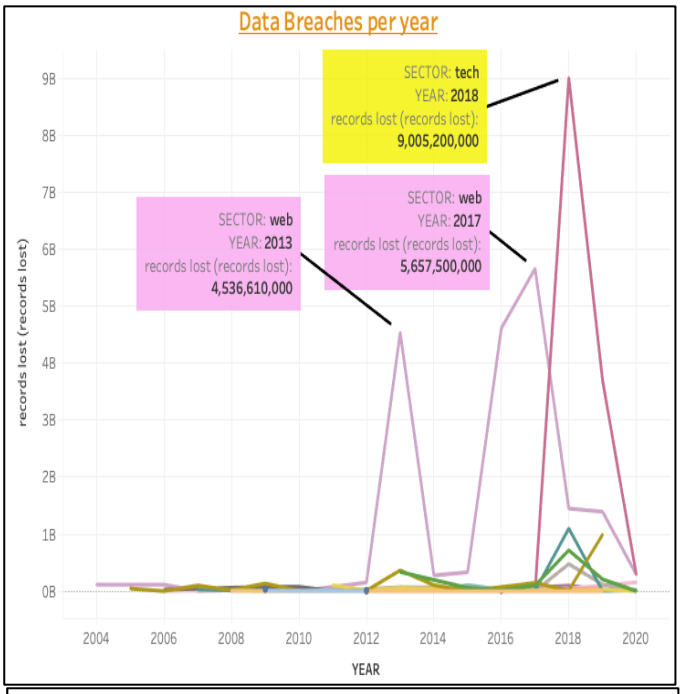


3) What Year from 2004 - 2020 did the most data breaches occur?

Answer:

I used line graphs to represent change of data breaches over time. Line graphs are slightly better than bar graphs to visualise changes over time as smaller changes are more visually seen.

- We can see many different peaks which occur between 2004 and 2020. The highest peak belongs to 2018 with over 12 billion records lost as seen in fig. 13. The second peak belongs to 2013 with over 5 billion records lost in that year, following a strong dip in 2014, then increasing in 2015.
- Next we are able to compare different sectors to analyse the records lost per year. We can see that the Tech industry is the major cause of the highest peak in 2018 with over 9 billion leaks.
- The Web industry has also contributed to majority of the leaks in 2017 as well as 2013.
- The lowest peak was found in 2005 with just over 44 million records lost.
- We can see a sudden increase in data breaches in 2018 as compared to 2013. Furthermore, we can see an exponential increase in data breaches from 2005 to 2018.



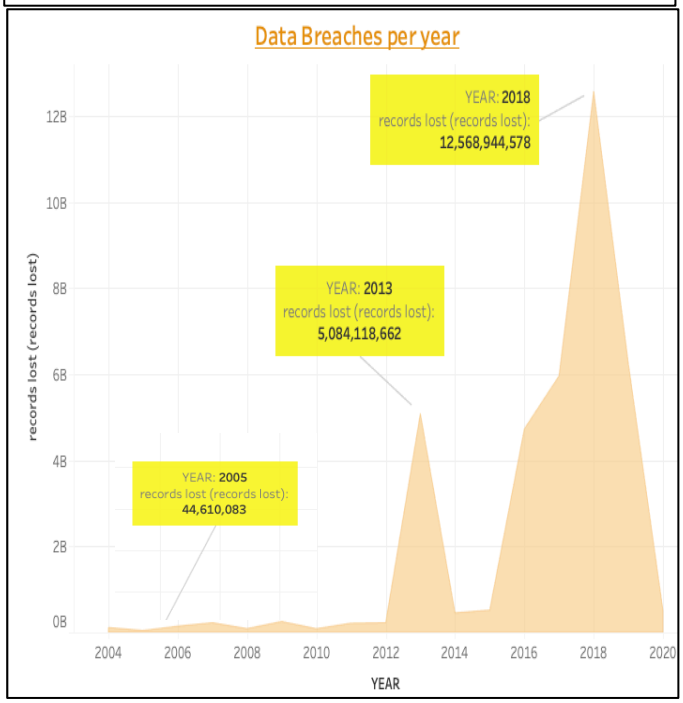


Figure 12.(top) Data breaches per year- with each sector

Figure 13. (bottom)Data breaches per year- overview

4) How sensitive the data is that's being lost?

Answer:
The data sensitivity defines the level of private information being leaked; eg. Very High sensitivity means full personal details, very low means only email address being leaked.

- The graphs below show relationship of sensitivity level and records lost
- Figure 15 indicates, the level of sensitivity being leaked during 2004-2020 has been "Very Low" with just over 13 billion leaks of email address.
- "Very high" level has been the least over the years with less than 1 billion leaks of very high sensitivity data.
- Figure 14. Shows the Tech and Web industry being the highest in data leaks for "very low" sensitivity.
- Tech and web industry/sector would have the most leaks due to the large amount of technologies and companies moving online. Thus, more people are adding their personal data online.

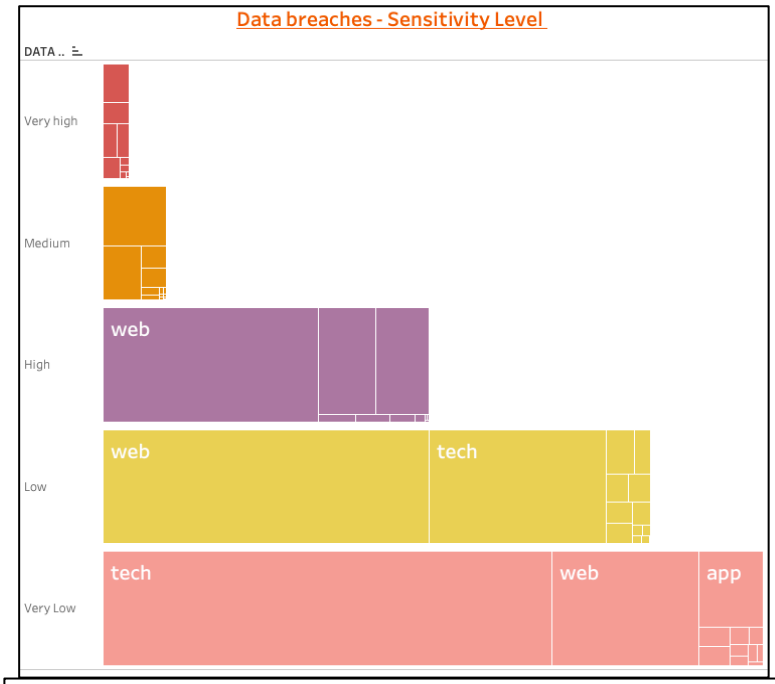


Figure 14. (On the left). Treemap bar graph of data sensitivity level, with division of each sector.

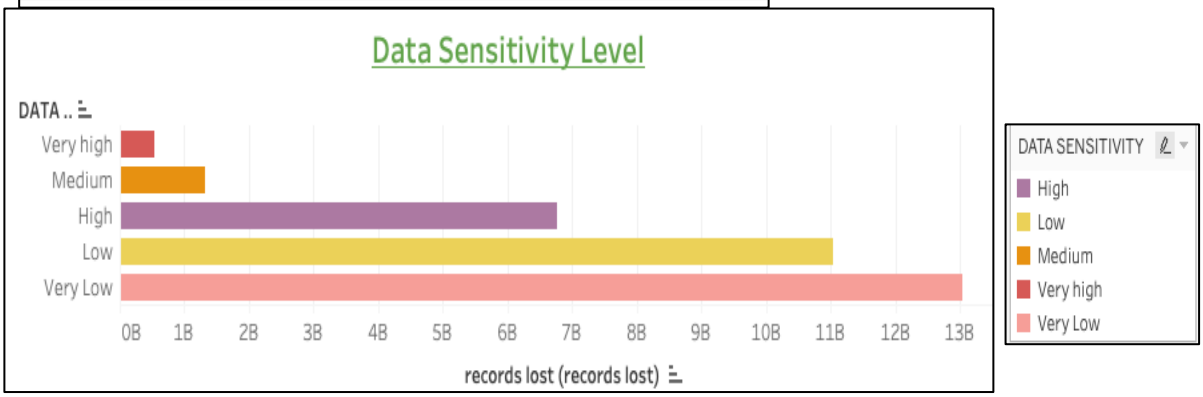
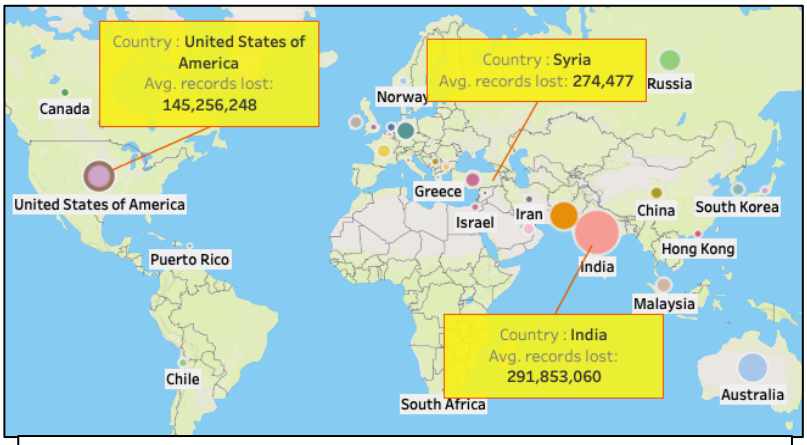


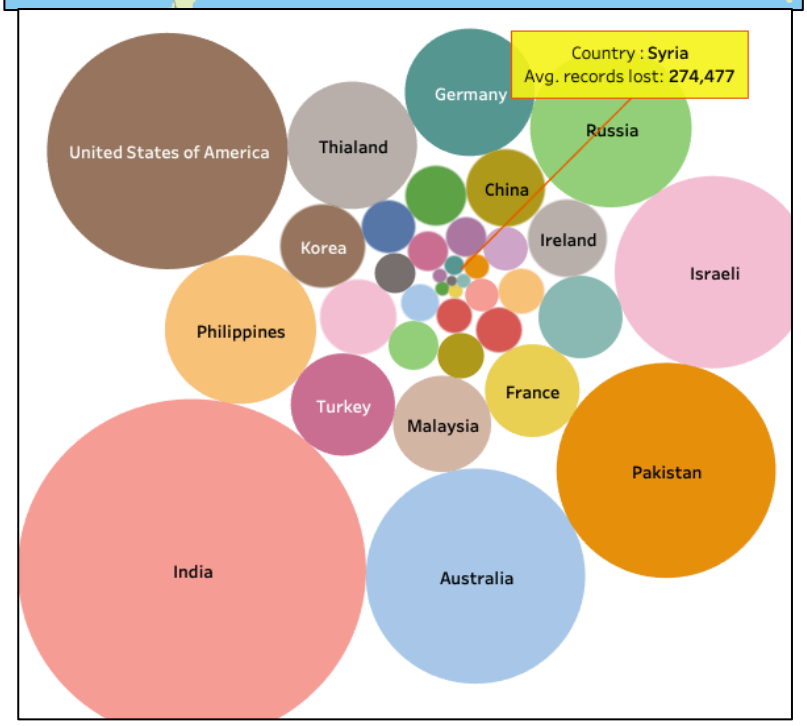
Figure. 15. (Below). Bar graph of data sensitivity level.

5) What country has the most data breaches?

Answer:
NOTE: Collecting this data was done manually;
I had gone through every company/ entity manually and found its location to find what country has the most data breaches, I had then created a data frame and had joined the data from the original file to collect data for records lost.
I collected the average of the records lost for each country to get an accurate result.

- From the graph in fig 16. we can see a small cluster of data breaches which occur around Europe.
- We can also see Syria being responsible for the least amount of breaches with 200 thousand breaches.
- India has the most data breaches with over 2 million breaches.
- From our collected data I am able to make a bubble chart as seen In figure 17. We are able to visually see, which country has the highest breach; India, United states, Pakistan.





- I created a map to analyse spatial data, this method enables me to see which points/ countries data breaches have occurred.
- The size of the points enable us to see the proportion of average record lost.

Figure. 16. (Above) World map o Data breaches.
Figure. 17 (Below)- Bubble chart Proportion of country with avg. records lost.

5. Conclusion

Overall we can see that the most common reason for data breaches to occur was, poor security and hacks The majority of this occurred in sectors such as the tech industry and the web sector. This indicates the necessity of having data regulations with strict data security laws around the tech and web sector as they're prone to major data attacks. The level of sensitivity is low with the tech industry contributing to the most data breaches. The most data attacks are found in India and the United States.

This **process of data exploration** of Data breaches with different sectors amongst different countries was successful, as I was able to answer all my questions. I was able to correctly visualise my data through the use to Tableau and excel, I could also find errors withing the raw data and clean them. This helped getting an accurate display of my data. I had made **transformation** of my data by **extracting** columns from my original file and creating new data frames. I also took the time to create a new data frame which gave me locations of each company, this further **enhanced** my exploration. I was not only able to answer my initial questions, but was also add and answer additional questions. I was able to see different patterns and trends within the different visualisations.

6. Reflection

This project gave me an opportunity to explore and wrangle data. I have gained a thorough understanding In the importance of the visualisation of data. I have also learned that checking data and cleaning data are curtail elements in creating accurate graphs. I realised that finding the correct data source is hard to find and data wrangling takes up most of the time. The questions have helped me understand how to analyse and create different graphs around different questions and that theirs multiple ways to represent data. This project has helped me improve my Tableau knowledge and has given me insight into the data science world.

7. Bibliography

David M, Tom E, Paul B, Dr Stephanie S, Duncan G (2020, May 11). World's Biggest Breaches and Hacks. Retrieved from. https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Paul B (2020, March, 3) . Which countries have the worst and best cyber security? Retrieved from. https://www.comparitech.com/blog/vpn-privacy/cybersecurity-by-country/