SCSE20-0940

Open source intelligence gathering and analysis of correlation of cyber risks

Submitted by: Aloysious Foo
Matriculation Number: U1822019J
Supervisor: Assoc Prof Anwitaman Datta

School of Computer Science and Engineering

A final year project report presented to Nanyang Technological University in partial fulfilment of the requirements for the degree of Bachelor of Engineering (Computer Science)

2021

Table of Contents

Abstract

Cyber attacks are happening every day, with many being oblivious to it and only start to realise the seriousness of it when it happens to them and by the time it will be too late. The impact of cyber threats can be potentially disastrous thus we need to be self educated on them so that we can be better prepared to defend against it with minimal impacts as far as possible. In order to achieve cyber knowledge, proper intelligence needs to be gathered to evaluate the attacks and understand what is going on and to possibly predict what these attacks will do in the future. This project aims to investigate the trends of cyber attacks, such as the types of attacks, who and what they target and their objective. It will be achieved by gathering openly available sources cyber incident data and analysis will be carried out to understand more on the data such as its distributions and their varieties.

Acknowledgement

I would like to take this opportunity to thank my project supervisor, Associate Professor Anwitaman Datta for his tips, guidance and continuous feedback throughout the time of the project till completion, which helped me to gain valuable experience and beneficial skills gained from the project.

List of Figures

Introduction

Cyber attacks are malicious activities launched by people through unauthorized access to computers illegally with intentions to steal private information or to gain money out of it. It has become a prominent threat today with the evolution of technology especially with the development of the internet. Cyber threats are dangerous as they can cause confidential data to be stolen and even cause business entities to lose millions of dollars. Over the past two decades, the Internet Crime Center reported a steady increase of monetary damage caused by cyber crimes [1]. Attack occurrences have also increased over the years and the need for cyber security has become a necessity today. One example is the increase in the number of data breaches annually over the past 15 years in the USA [2]. These threats cannot be left unhandled and be dominant on the internet, therefore there is a need to understand what the idea behind these threats in order is to know how to fight them and prevent future damages.

In this project, we will be looking out for openly available cyber incident data repositories to investigate what are the common kinds of threats, what they do, what are the impacts caused by them and who are the victims of the threats. This is a crucial point to figure out as people can then know what to focus on exactly to protect themselves against. Cyber attacks are broad and wide in variety and if people do not know what are the specific threats, time and resources will be wasted trying to figure out something that is not helpful to know and resolve. As cyber attacks can occur everywhere, by gaining intelligence on the common behaviors and impacts of these threats, we can all be educated on the importance of cyber security and make the digital world a safer place to live and work on.

Implementation

This section explains the process of intelligence gathering all the way to data visualisation. The first step of the project is to gather intelligence of relevant topics by sourcing for openly available cyber incident datasets and repositories online. Expected information to be obtained from these sources will be a brief summary of the incident, date of incident, type of attack occurred, type of organisation of the victim, victim demographics and the impacts. Data can be pre prepared in compiled excel, xml, json formats, or they have to be manually scraped.

1. Data Acquisition

This project harvest data from these following sources:

No.	Source Name	File Type Available	Description of Source
1	Hackmageddon [3]	CSV	Hackmageddon is a free and open data repository on cyber incidents that is continuously updating its collection till present day.
2	Veris Community Database (VCDB) [4]	CSV	Vocabulary for Event Recording and Incident Sharing (VERIS) is a free available data repository that collects security incidents and shares

			with others and helps everyone to learn from experiences to better measure and manage risk.
3	Vaibhavi Awghate’s github repository [5]	CSV	This is a github profile of a data analyst who posted various coding projects on her repository, including data analysis projects in python.

2. Data Preprocessing and Analysis

The next step after acquiring data is to preprocess the data. Very often datasets are obtained raw without any modifications and there are many times that there will be inconsistencies, missing values and unnecessary information that is not needed for analysis hence it can be sliced off from the dataset to make it look neater and avoid confusion.

Once the dataset is cleaned and prepared in the way we want, it will be ready to perform analysis and visualisation to meet the objective. The objective of the analysis is to identify trends of the 3 different datasets and to possibly see common insights about cyber threats. The following stated are some of the visualisation tools and methods used for analysing the data using different python libraries(Pandas, Seaborn, Scikit-Learn):

- Bar Chart:
    - A bar chart is a chart that presents categorical data with rectangular bars vertically or horizontally with lengths proportional to the quantity of

values that they represent. It is used to show “how many” for each category [6].

- Pie Chart
  - A pie chart is a circular statistical chart divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents. It is used to show percentages of a whole [7].

- Countplot
  - A countplot serves the same function as a bar chart, to show “how many” in a particular category. The only difference is countplot is under the python visualisation library of Seaborn which offers more powerful options to visualise data compared to the regular Matplotlib [8].

- Scatterplot
  - A scatter plot is a plot using Cartesian coordinates to display pairs of numerical data, with one variable on each axis, to look for a relationship between them [9].

- K-nearest neighbor Regression
  - KNN regression is a supervised machine learning algorithm that predicts a continuous datapoint by referencing its nearest ‘k’ data points defined and averages their value to predict a value. It will be implemented from the scikit-learn library [10].

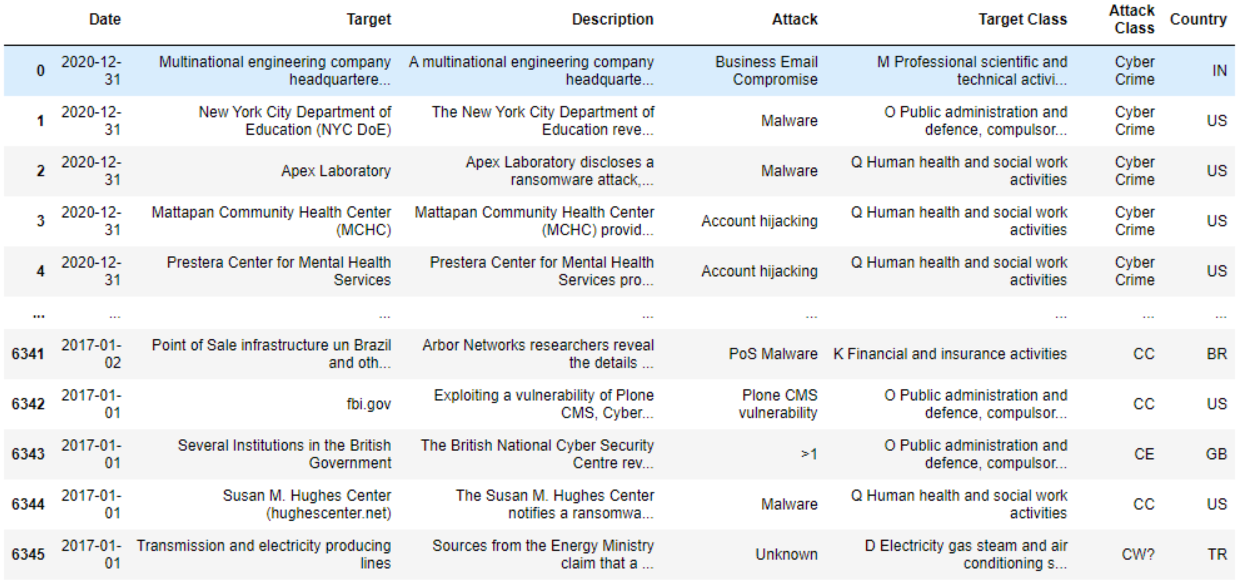
- Decision Tree Classifier
  - A decision tree is a supervised machine learning algorithm that uses a tree-like model that enables a decision about some kind of process to be made. A decision tree regressor predicts continuous data while a decision tree classifier predicts a category where the data should be under. Each branch

    - represents the outcome of the test, and each leaf node represents a class label [11].

- NLTK Pos Tagging
    - Part-of-speech (POS) tagging is a popular Natural Language Processing process that categorises each word in a text (corpus) in correspondence with a particular part of speech, such as noun, verbs, adjectives etc. The library that will be used is the Natural Language Toolkit(NLTK) in python [12].

2.1. Hackmageddon dataset

The first dataset we will analyze is the Hackmageddon dataset. This is a brief overview of the raw data of Hackmageddon. This dataset shows mostly categorical and descriptive data curated from 2017 to 2020.



	Date	Target	Description	Attack	Target Class	Attack Class	Country
0	2020-12-31	Multinational engineering company headquartere...	A multinational engineering company headquarte...	Business Email Compromise	M Professional scientific and technical activi...	Cyber Crime	IN
1	2020-12-31	New York City Department of Education (NYC DoE)	The New York City Department of Education reve...	Malware	O Public administration and defence, compulsor...	Cyber Crime	US
2	2020-12-31	Apex Laboratory	Apex Laboratory discloses a ransomware attack,...	Malware	Q Human health and social work activities	Cyber Crime	US
3	2020-12-31	Mattapan Community Health Center (MCHC)	Mattapan Community Health Center (MCHC) provid...	Account hijacking	Q Human health and social work activities	Cyber Crime	US
4	2020-12-31	Prestera Center for Mental Health Services	Prestera Center for Mental Health Services pro...	Account hijacking	Q Human health and social work activities	Cyber Crime	US
...	...	...	...	...	...	...	...
6341	2017-01-02	Point of Sale infrastructure un Brazil and oth...	Arbor Networks researchers reveal the details ...	PoS Malware	K Financial and insurance activities	CC	BR
6342	2017-01-01	fbi.gov	Exploiting a vulnerability of Plone CMS, Cyber...	Plone CMS vulnerability	O Public administration and defence, compulsor...	CC	US
6343	2017-01-01	Several Institutions in the British Government	The British National Cyber Security Centre rev...	>1	O Public administration and defence, compulsor...	CE	GB
6344	2017-01-01	Susan M. Hughes Center (hughescenter.net)	The Susan M. Hughes Center notifies a ransomwa...	Malware	Q Human health and social work activities	CC	US
6345	2017-01-01	Transmission and electricity producing lines	Sources from the Energy Ministry claim that a ...	Unknown	D Electricity gas steam and air conditioning s...	CW?	TR

Figure 1: Hackmageddon dataset

2.1.1. Preprocessing

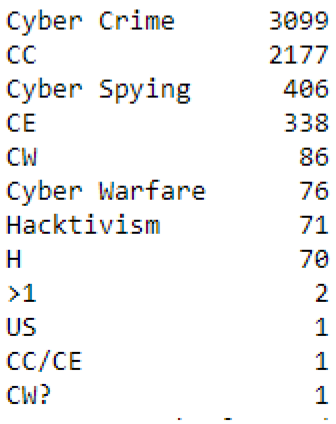
One of the data inconsistencies is that there are unclear labels and duplicates which are named as something else similar under some columns such as the ‘Attack Class’, ‘Target’ and ‘Attack’ as shown.

Attack Class:

There are 4 types of attack: Cyber Crime, Cyber Spying, Cyber Warfare and Hacktivism. Anything else that is not clearly defined will be dropped and duplicate attack types will be combined.

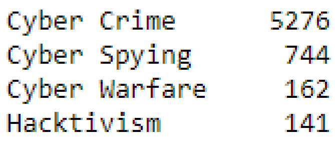
Before

After



```
Cyber Crime      3099
CC               2177
Cyber Spying      406
CE                338
CW                 86
Cyber Warfare      76
Hacktivism         71
H                  70
>1                  2
US                  1
CC/CE               1
CW?                 1
```

Figure 2: Attack Class with duplicate



```
Cyber Crime      5276
Cyber Spying      744
Cyber Warfare     162
Hacktivism        141
```

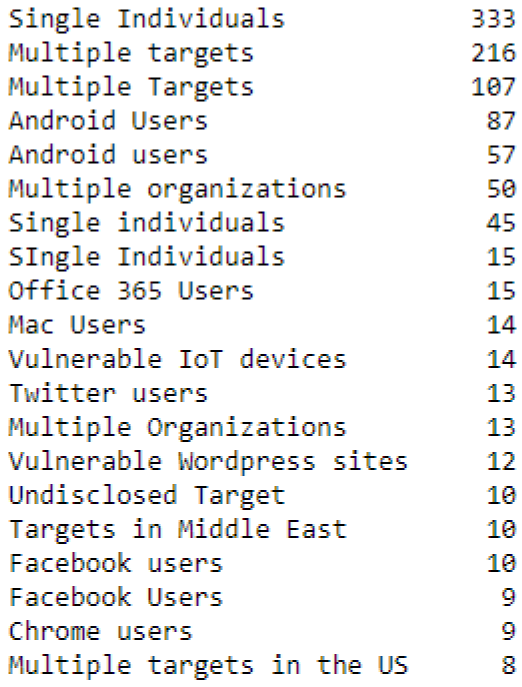
Figure 3: Attack Class without duplicate

Target:

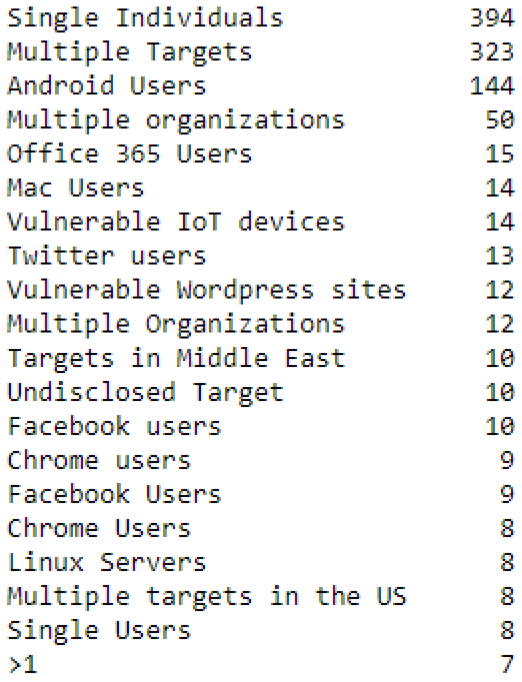
There are few duplicate values due to the capitalisation difference such as ‘targets’ and ‘Targets’ and will be combined.

Before

After



```
Single Individuals            333
Multiple targets              216
Multiple Targets              107
Android Users                  87
Android users                  57
Multiple organizations         50
Single individuals             45
SIngle Individuals             15
Office 365 Users               15
Mac Users                      14
Vulnerable IoT devices         14
Twitter users                  13
Multiple Organizations         13
Vulnerable Wordpress sites     12
Undisclosed Target             10
Targets in Middle East         10
Facebook users                 10
Facebook Users                  9
Chrome users                    9
Multiple targets in the US      8
```



```
Single Individuals            394
Multiple Targets              323
Android Users                 144
Multiple organizations         50
Office 365 Users               15
Mac Users                      14
Vulnerable IoT devices         14
Twitter users                  13
Vulnerable Wordpress sites     12
Multiple Organizations         12
Targets in Middle East         10
Undisclosed Target             10
Facebook users                 10
Chrome users                    9
Facebook Users                  9
Chrome Users                    8
Linux Servers                   8
Multiple targets in the US      8
Single Users                    8
>1                              7
```

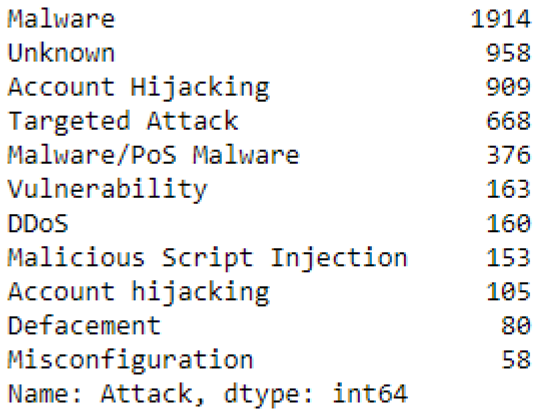
Figure 4: Target Class with duplicate   *Figure 5: Target Class without duplicate*

Attack:

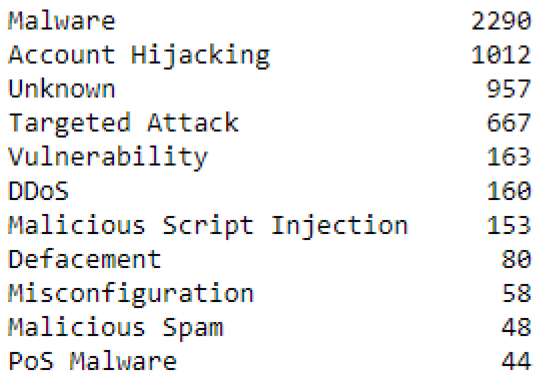
The duplicates found are Malware and Account Hijacking and will be combined.

Before

After



```
Malware                       1914
Unknown                        958
Account Hijacking              909
Targeted Attack                668
Malware/PoS Malware            376
Vulnerability                  163
DDoS                           160
Malicious Script Injection     153
Account hijacking              105
Defacement                      80
Misconfiguration                58
Name: Attack, dtype: int64
```



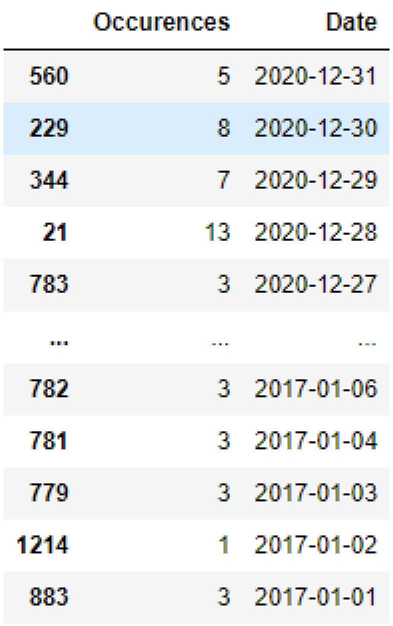
```
Malware                       2290
Account Hijacking             1012
Unknown                        957
Targeted Attack                667
Vulnerability                  163
DDoS                           160
Malicious Script Injection     153
Defacement                      80
Misconfiguration                58
Malicious Spam                  48
PoS Malware                     44
```

Figure 6: Attack type with duplicate   *Figure 7: Attack type without duplicate*

2.1.2. Analysis

2.1.2.1. Attack Occurrences

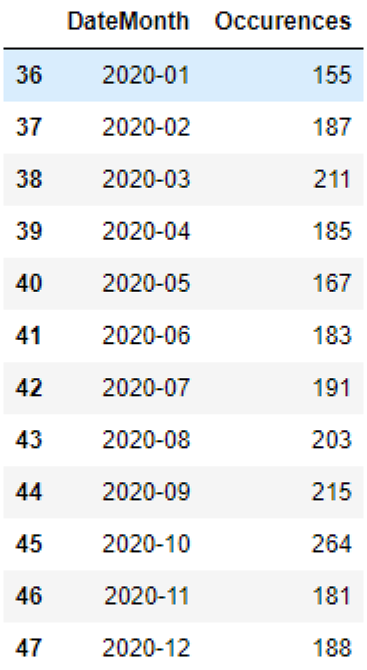
The ‘Date’ column is processed to find out the number of attack occurrences per day as shown.



	Occurences	Date
560	5	2020-12-31
229	8	2020-12-30
344	7	2020-12-29
21	13	2020-12-28
783	3	2020-12-27
...	...	...
782	3	2017-01-06
781	3	2017-01-04
779	3	2017-01-03
1214	1	2017-01-02
883	3	2017-01-01

Figure 8: Attack occurrence daily

It is further processed to calculate occurrences per month to better visualise it later.



	DateMonth	Occurences
36	2020-01	155
37	2020-02	187
38	2020-03	211
39	2020-04	185
40	2020-05	167
41	2020-06	183
42	2020-07	191
43	2020-08	203
44	2020-09	215
45	2020-10	264
46	2020-11	181
47	2020-12	188

Figure 9: Attack occurrence monthly

2.1.2.2. Trend analysis from ‘Attack Types’, ‘Target Class’

The major features ‘Attack Types’, ‘Target Class’ are also analysed. From here on, we will observe the significance of the data which is to find out what the most frequent attack is and the victim classes that were targeted most by it . We will then try to find out the exact targets as close as possible by continuously analysing the data deeper.

2.1.2.3. Action words from incident description

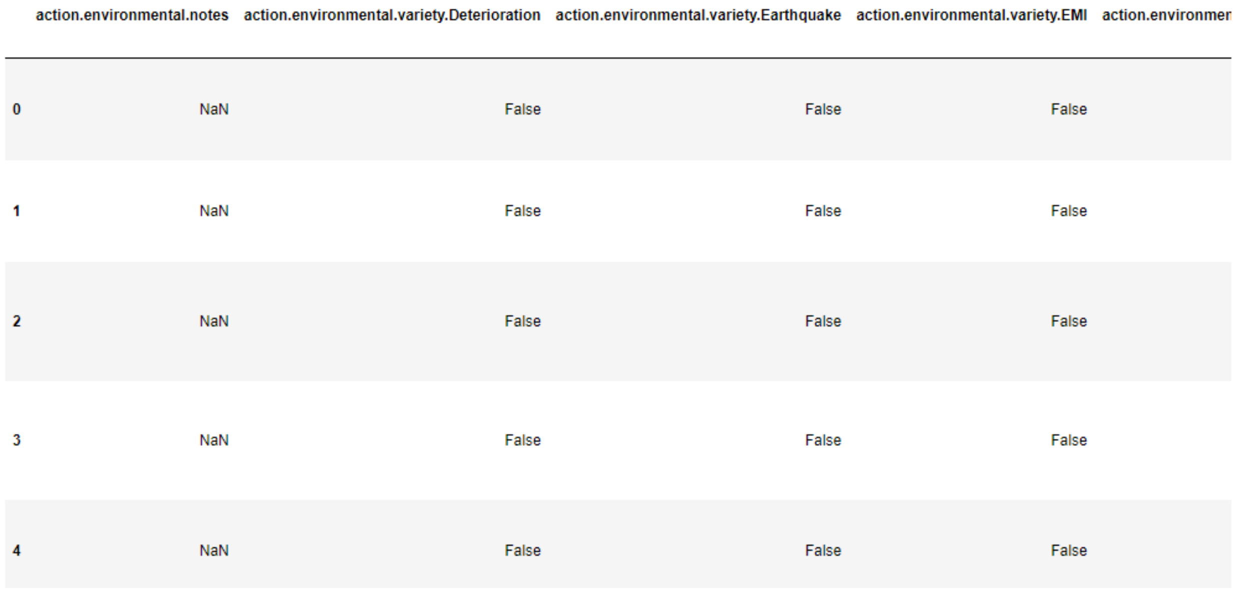
Verbs will be extracted under the incident description column using nltk’s pos tagging to briefly look at what are the common actions done from a cyber attack. These words will be identified based on the cyber context.

2.1.2.4. Trend analysis from Country

Another significance to analyse is the country that gets hit by cyber attack the most. From there on, we will also visualise the frequent attack types and victims classes targeted in that particular country and use Named Entity Recognition to find out as exact as possible what the attacks are aiming.

2.2. Veris Community Database dataset

The 2nd dataset is the VCDB dataset. Unlike normal datasets that label the column name and give different categories, this dataset contains every possible category labelled as a column with the value either TRUE or FALSE. Information cannot be extracted in a straightforward manner and require manipulation.



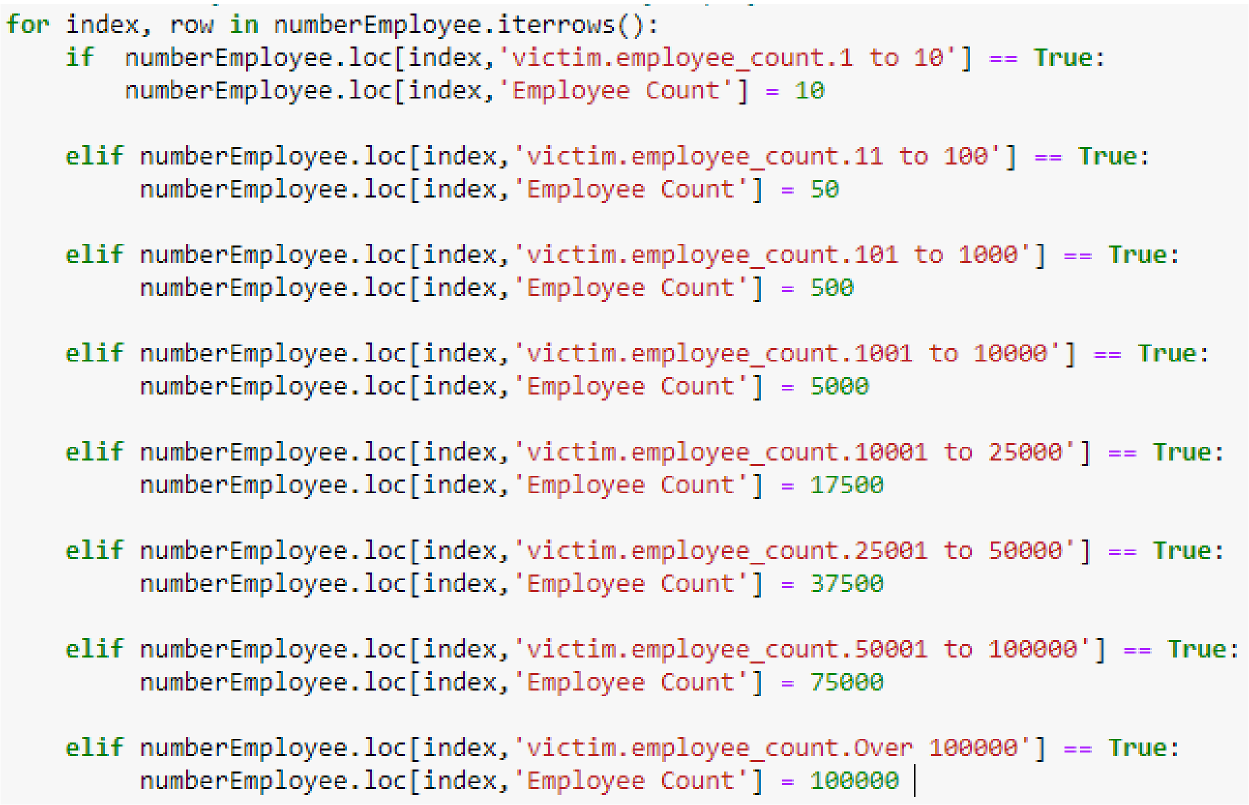
	action.environmental.notes	action.environmental.variety.Deterioration	action.environmental.variety.Earthquake	action.environmental.variety.EMI	action.environmer
0	NaN	False	False	False	
1	NaN	False	False	False	
2	NaN	False	False	False	
3	NaN	False	False	False	
4	NaN	False	False	False	

Figure 10: *Veris dataset*

2.2.1. Preprocessing

6 attributes will be selected for analysis:
Incident summary, Employee count, Revenue, Incident Type, Year of incident, Impact loss.

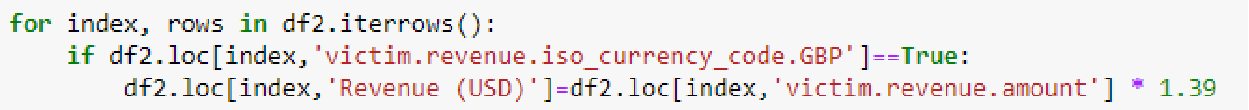
For Employee count, the values are not discrete but given in 8 kinds: 1-10, 11-100, 101-1000, 1001-10000, 10001-25000, 25001-50000, 50001-100000 and more than 100000. Since the exact number is unknown, we will take the median of each range: 10, 50, 500, 5000, 17500, 37500, 75000 and 100000 and store it in a new column ‘Employee Count’.



```
for index, row in numberEmployee.iterrows():
    if  numberEmployee.loc[index,'victim.employee_count.1 to 10'] == True:
        numberEmployee.loc[index,'Employee Count'] = 10

    elif numberEmployee.loc[index,'victim.employee_count.11 to 100'] == True:
         numberEmployee.loc[index,'Employee Count'] = 50

    elif numberEmployee.loc[index,'victim.employee_count.101 to 1000'] == True:
         numberEmployee.loc[index,'Employee Count'] = 500

    elif numberEmployee.loc[index,'victim.employee_count.1001 to 10000'] == True:
         numberEmployee.loc[index,'Employee Count'] = 5000

    elif numberEmployee.loc[index,'victim.employee_count.10001 to 25000'] == True:
         numberEmployee.loc[index,'Employee Count'] = 17500

    elif numberEmployee.loc[index,'victim.employee_count.25001 to 50000'] == True:
         numberEmployee.loc[index,'Employee Count'] = 37500

    elif numberEmployee.loc[index,'victim.employee_count.50001 to 100000'] == True:
         numberEmployee.loc[index,'Employee Count'] = 75000

    elif numberEmployee.loc[index,'victim.employee_count.Over 100000'] == True:
         numberEmployee.loc[index,'Employee Count'] = 100000
```

Figure 11: Assign Employee Count

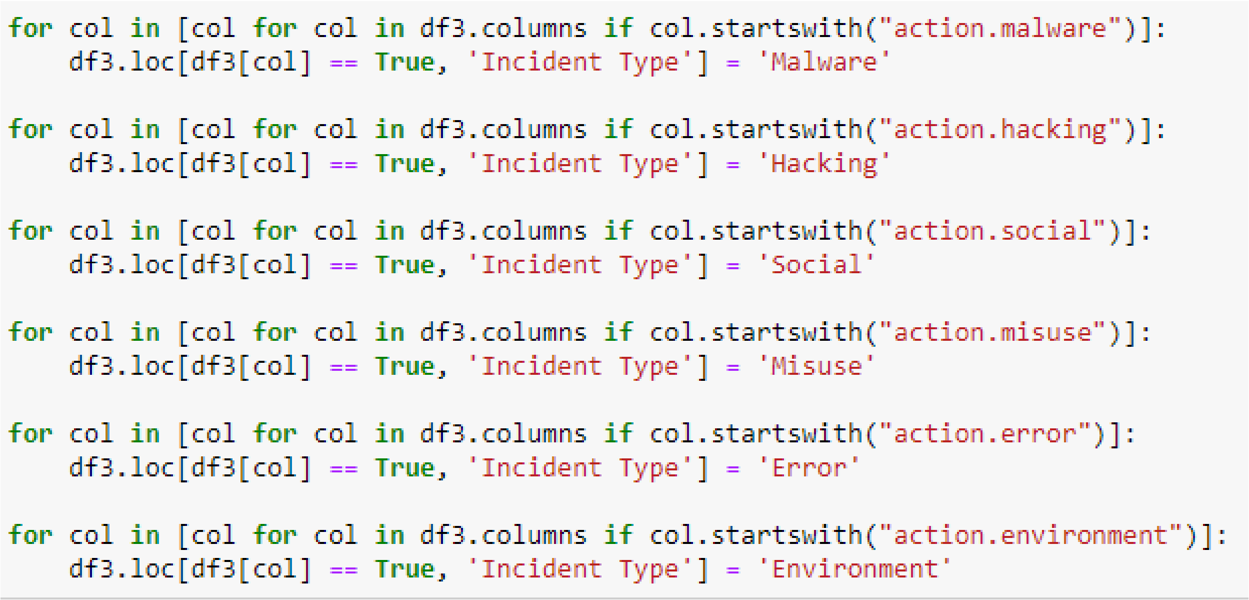
For the company’s revenue, after filtering off the missing values there are 505 entries with only USD and GBP currency. The currency has been standardised to USD and all entries with GBP have been converted to USD and stored in a new column ‘Revenue (USD)’ as shown.



```
for index, rows in df2.iterrows():
    if df2.loc[index,'victim.revenue.iso_currency_code.GBP']==True:
        df2.loc[index,'Revenue (USD)']=df2.loc[index,'victim.revenue.amount'] * 1.39
```

Figure 12: Convert GBP to USD

For incident types there are malware, hacking, social, misuse, error and environment types. Within these 6 types there are also subtypes(e.g backdoor, adware, brute force under malware) but we do not need to know the subtypes so we can extract the rows and indicate the main type in a new column ‘Incident Type’ as long as any of the subtypes is TRUE.

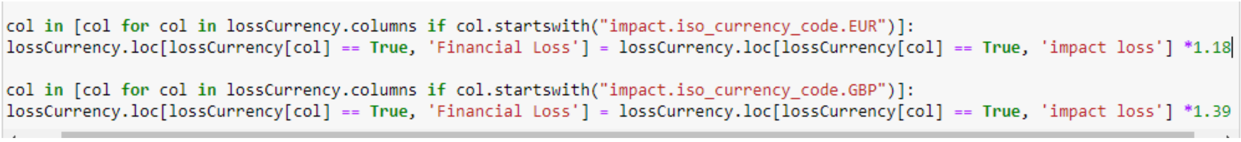


```
for col in [col for col in df3.columns if col.startswith("action.malware")]:
    df3.loc[df3[col] == True, 'Incident Type'] = 'Malware'

for col in [col for col in df3.columns if col.startswith("action.hacking")]:
    df3.loc[df3[col] == True, 'Incident Type'] = 'Hacking'

for col in [col for col in df3.columns if col.startswith("action.social")]:
    df3.loc[df3[col] == True, 'Incident Type'] = 'Social'

for col in [col for col in df3.columns if col.startswith("action.misuse")]:
    df3.loc[df3[col] == True, 'Incident Type'] = 'Misuse'

for col in [col for col in df3.columns if col.startswith("action.error")]:
    df3.loc[df3[col] == True, 'Incident Type'] = 'Error'

for col in [col for col in df3.columns if col.startswith("action.environment")]:
    df3.loc[df3[col] == True, 'Incident Type'] = 'Environment'
```

Figure 13: Assign Incident Type

For impact loss, the available currencies recorded after filtering off missing values are EUR and GBP and will be standardised to USD. A new column ‘Financial Loss’ is copied from the original ‘impact loss’ column and ‘Financial Loss’ will be processed and recomputed to the amount in USD.

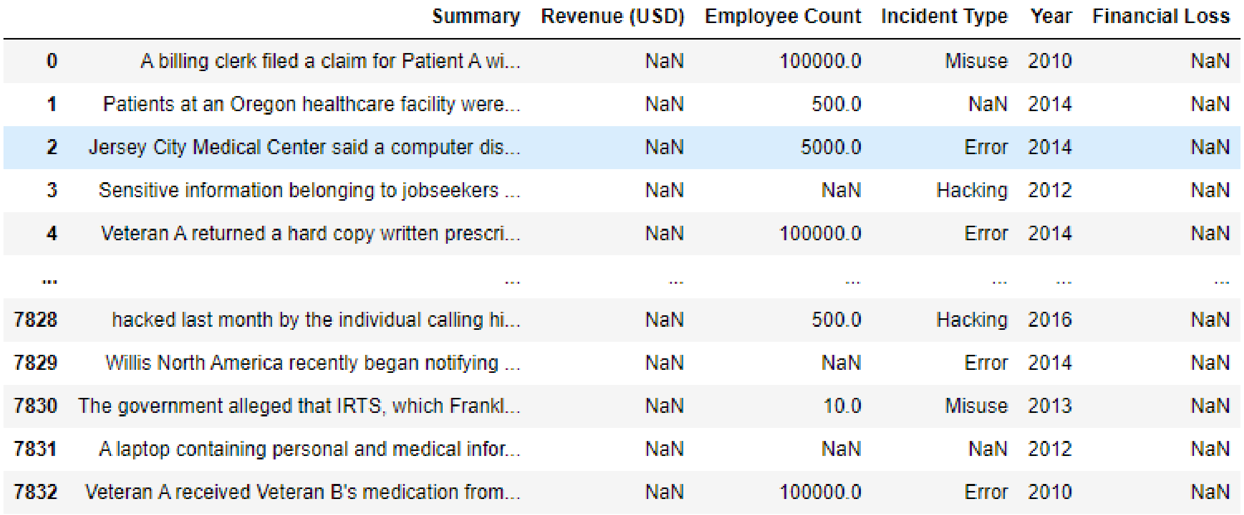
Figure 14: Convert all revenue currency to USD



```
col in [col for col in lossCurrency.columns if col.startswith("impact.iso_currency_code.EUR")]:
lossCurrency.loc[lossCurrency[col] == True, 'Financial Loss'] = lossCurrency.loc[lossCurrency[col] == True, 'impact loss'] *1.18

col in [col for col in lossCurrency.columns if col.startswith("impact.iso_currency_code.GBP")]:
lossCurrency.loc[lossCurrency[col] == True, 'Financial Loss'] = lossCurrency.loc[lossCurrency[col] == True, 'impact loss'] *1.39
```

2.2.2. Analysis

Now that the relevant information has been processed in a neater format, it is ready for visualisation.



	Summary	Revenue (USD)	Employee Count	Incident Type	Year	Financial Loss
0	A billing clerk filed a claim for Patient A wi...	NaN	100000.0	Misuse	2010	NaN
1	Patients at an Oregon healthcare facility were...	NaN	500.0	NaN	2014	NaN
2	Jersey City Medical Center said a computer dis...	NaN	5000.0	Error	2014	NaN
3	Sensitive information belonging to jobseekers ...	NaN	NaN	Hacking	2012	NaN
4	Veteran A returned a hard copy written prescri...	NaN	100000.0	Error	2014	NaN
...	...	...	...	...	...	...
7828	hacked last month by the individual calling hi...	NaN	500.0	Hacking	2016	NaN
7829	Willis North America recently began notifying ...	NaN	NaN	Error	2014	NaN
7830	The government alleged that IRTS, which Frankl...	NaN	10.0	Misuse	2013	NaN
7831	A laptop containing personal and medical infor...	NaN	NaN	NaN	2012	NaN
7832	Veteran A received Veteran B's medication from...	NaN	100000.0	Error	2010	NaN

Figure 15: Veris cleaned dataset

2.2.2.1. Categorical data distributions

The incident year, incident type and average employee count are being visualised to see the number of types of occurrences.

2.2.2.2. Financial loss to categorical data

The financial loss is visualised to see how each of these attributes, the incident type, average employee count and incident year will affect the losses.

2.2.2.3. Revenue to incident type

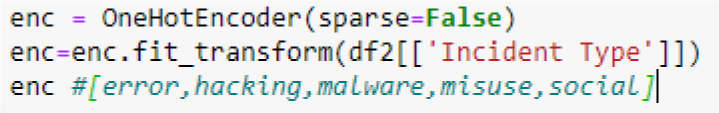
Visualisation is done to see how the amount of revenue determines the kind of attacks the company will face as this provides knowledge of whether attacks target low or high revenue companies.

2.2.2.4. Prediction of financial loss

Using the incident type, incident year and average employee count as predictors, we can try to predict the loss expected. This allows companies to know and have plans beforehand how much they are likely to lose if they were to be hit by a certain category of attack.

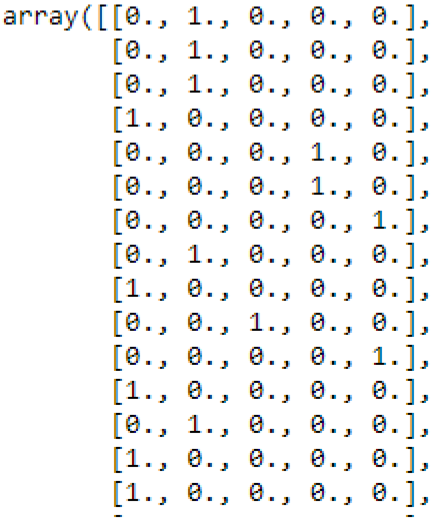
2.2.2.4.1. One-Hot Encoding

Machine learning only recognises numbers as its input and cannot recognise words hence non-numeric data must be encoded first. Categorical variables come under ordinal and nominal. Ordinal data has an element of order/ranking(e.g. high/med/low, good/very good/excellent) and nominal does not have element of ranking(e.g color, gender). The data that we have is under nominal thus we will use One-Hot Encoding to convert the data to numbers.One-Hot Encoding adds an array of binary values to the column values, indicating 0 for absent and 1 for present. For example, if we encode the incident type it will show an array of 5 binary values(e.g [0,1,0,0,0]) as there are only 5 incident types. This is achieved by doing what is shown below as an example.



```
enc = OneHotEncoder(sparse=False)
enc=enc.fit_transform(df2[['Incident Type']])
enc #[error,hacking,malware,misuse,social]
```

Figure 16: Fit and transform encoding



```
array([[0., 1., 0., 0., 0.],
       [0., 1., 0., 0., 0.],
       [0., 1., 0., 0., 0.],
       [1., 0., 0., 0., 0.],
       [0., 0., 0., 1., 0.],
       [0., 0., 0., 1., 0.],
       [0., 0., 0., 0., 1.],
       [0., 1., 0., 0., 0.],
       [1., 0., 0., 0., 0.],
       [0., 0., 1., 0., 0.],
       [0., 0., 0., 0., 1.],
       [1., 0., 0., 0., 0.],
       [0., 1., 0., 0., 0.],
       [1., 0., 0., 0., 0.],
       [1., 0., 0., 0., 0.],
```

Figure 17: One-Hot Encoding

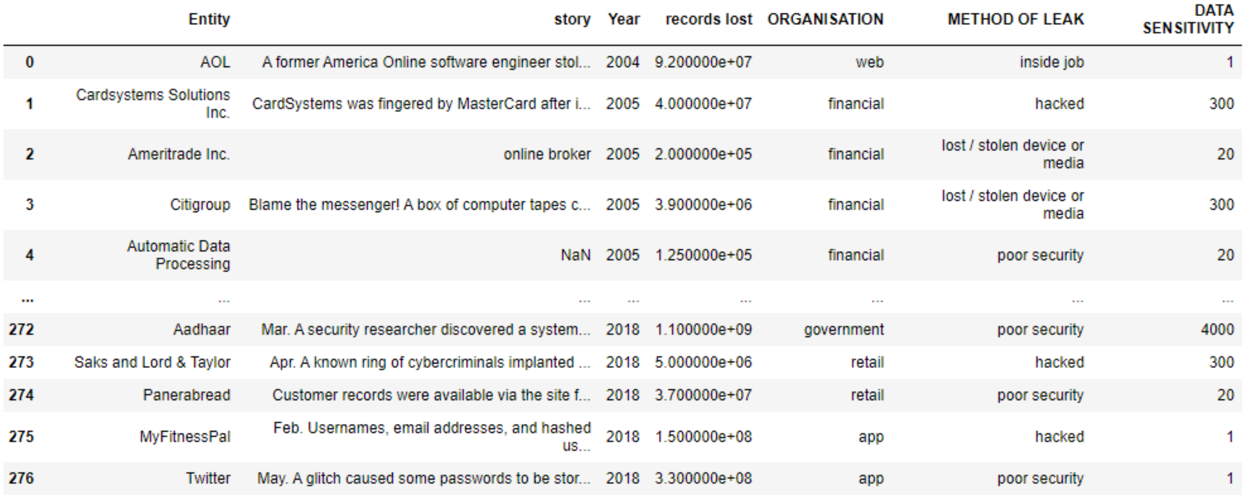
The results show that the incident type is' hacking' for row 1-3 and 'error' for row 4 and so on. After conversion, the data can then be input into the KNN Regression model to do prediction.

2.2.2.5. Predict Incident Type based on financial loss

Based on the amount of losses, a model is created to predict what will be the possible incident type. This provides knowledge of the probability of attack category given the amount of losses one has.

2.3. Github dataset

The 3rd dataset used is a collection of data breach incidents from a data analyst's github repository which is free and open. Information on the dataset is shown below.



	Entity	story	Year	records lost	ORGANISATION	METHOD OF LEAK	DATA SENSITIVITY
0	AOL	A former America Online software engineer stol...	2004	9.200000e+07	web	inside job	1
1	Cardsystems Solutions Inc.	CardSystems was fingered by MasterCard after i...	2005	4.000000e+07	financial	hacked	300
2	Ameritrade Inc.	online broker	2005	2.000000e+05	financial	lost / stolen device or media	20
3	Citigroup	Blame the messenger! A box of computer tapes c...	2005	3.900000e+06	financial	lost / stolen device or media	300
4	Automatic Data Processing	NaN	2005	1.250000e+05	financial	poor security	20
...	...	...	...	...	...	...	...
272	Aadhaar	Mar. A security researcher discovered a system...	2018	1.100000e+09	government	poor security	4000
273	Saks and Lord & Taylor	Apr. A known ring of cybercriminals implanted ...	2018	5.000000e+06	retail	hacked	300
274	Panerabread	Customer records were available via the site f...	2018	3.700000e+07	retail	poor security	20
275	MyFitnessPal	Feb. Usernames, email addresses, and hashed us...	2018	1.500000e+08	app	hacked	1
276	Twitter	May. A glitch caused some passwords to be stor...	2018	3.300000e+08	app	poor security	1

Figure 18: Github dataset

2.3.1. Analysis

2.3.1.1. Categorical data distributions

The incident year, method of leak and type of organisation count are being visualised to see the number of occurrences for each type.

2.3.1.2. Records lost to method of leak

The method of leak is visualised to see how it affects the relationship to the number of records lost.

2.3.1.3. Predict records lost based on method of leak

The column 'method of leak' is used to predict the records lost from a future data breach. As the method of leak is a categorical data type, it has to be encoded into numeric form first by doing One-Hot Encoding.

2.3.1.4. Predict type of organisation to be targeted based on year of attack

The type of organisation to be hit by a data breach is predicted based on the year of attack to see if it is possible to visualise whether there is a growing/change in trend of data breaches that targets specific types of organisations over the years.

Results

Evaluating results fulfills the entire purpose of data analysis because it draws important information about the dataset and allows the audience to make proper decisions in the future which are relevant to the practical issues discovered from analysing the dataset.

This section shows all output results generated from data visualisation and machine learning tools, including charts, graphs, word entity classifications, value predictions and accuracy scores. Inferences are drawn to gain insight about the information in the dataset.

1. Hackmageddon dataset:

1.1. Trends

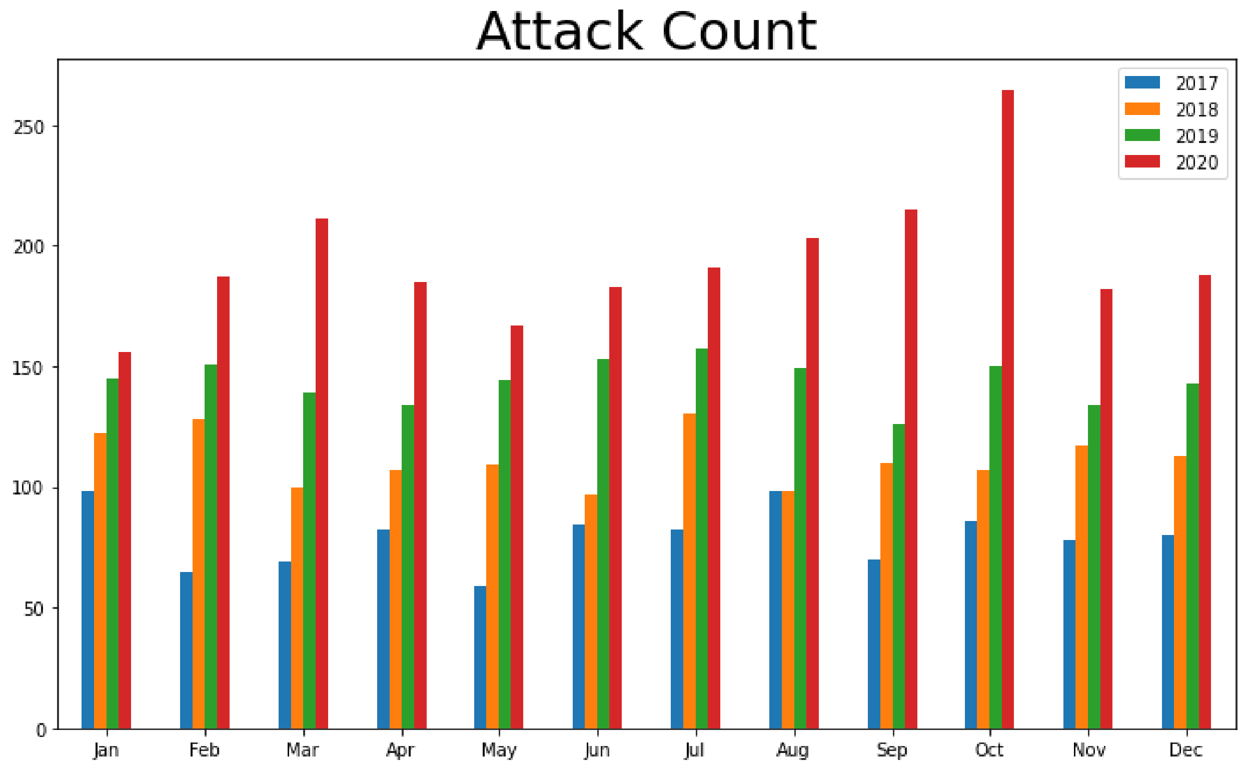


Figure 19: Attack occurrence 2017-2020

The graph shows the number of attacks in the dataset within a span of 4 years from 2017-2020. We can tell that cyber attacks are happening more often as time goes by and it is vital to bring cyber security into the topic as internet technologies are advancing over the years.

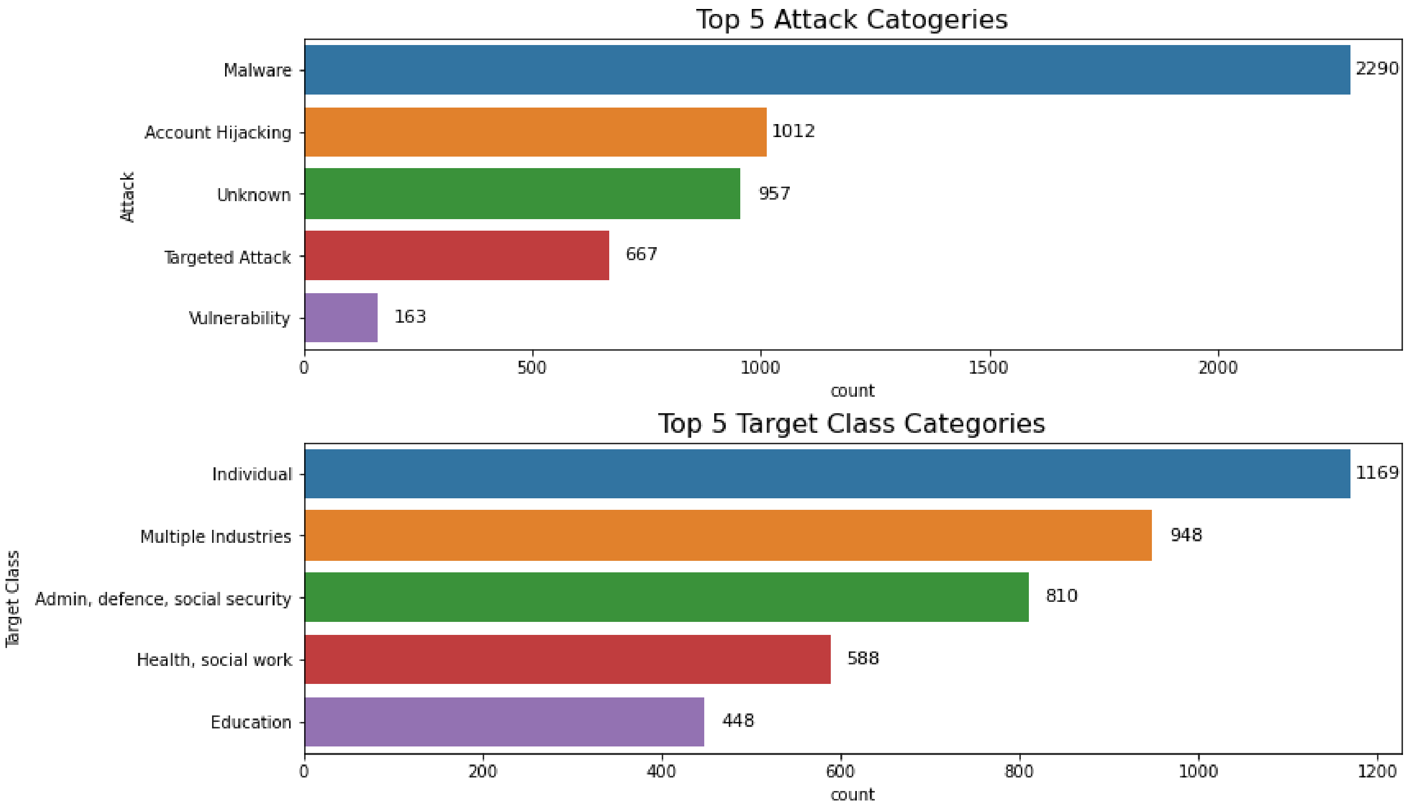


Figure 20: Attack and target class categories

Under the attack category, malware is the one that happens the most frequently. Under the target class category, we can tell that cyber attacks seem to be targeting individuals and industries the most, probably to steal personal information or money from them.

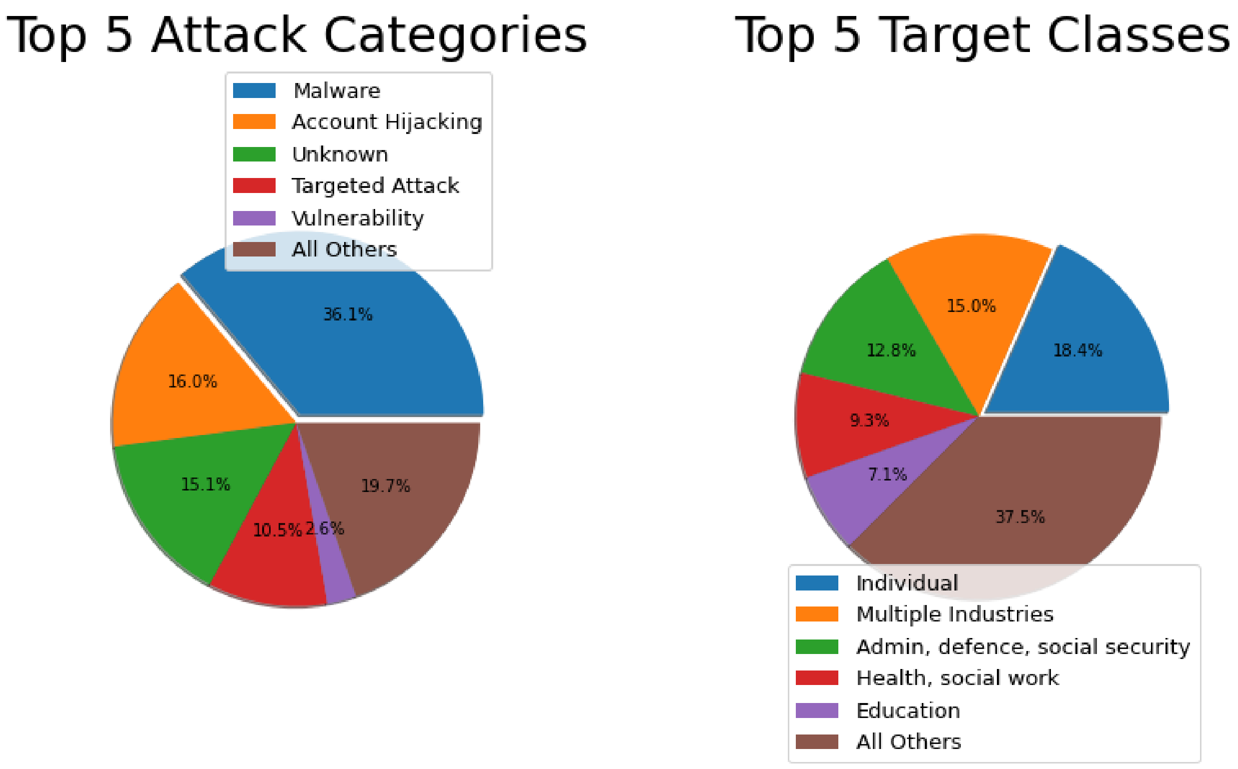


Figure 21: Attack and target class percentages

From the percentage distributions, it can be seen that malware is significant in being the main type of attack, outweighing all other types of attacks combined. This may be because malware can be optimised by consistent modification as software technology improves and malware can come with many kinds. It may be useful if we can further analyse to find out about malware from this dataset.

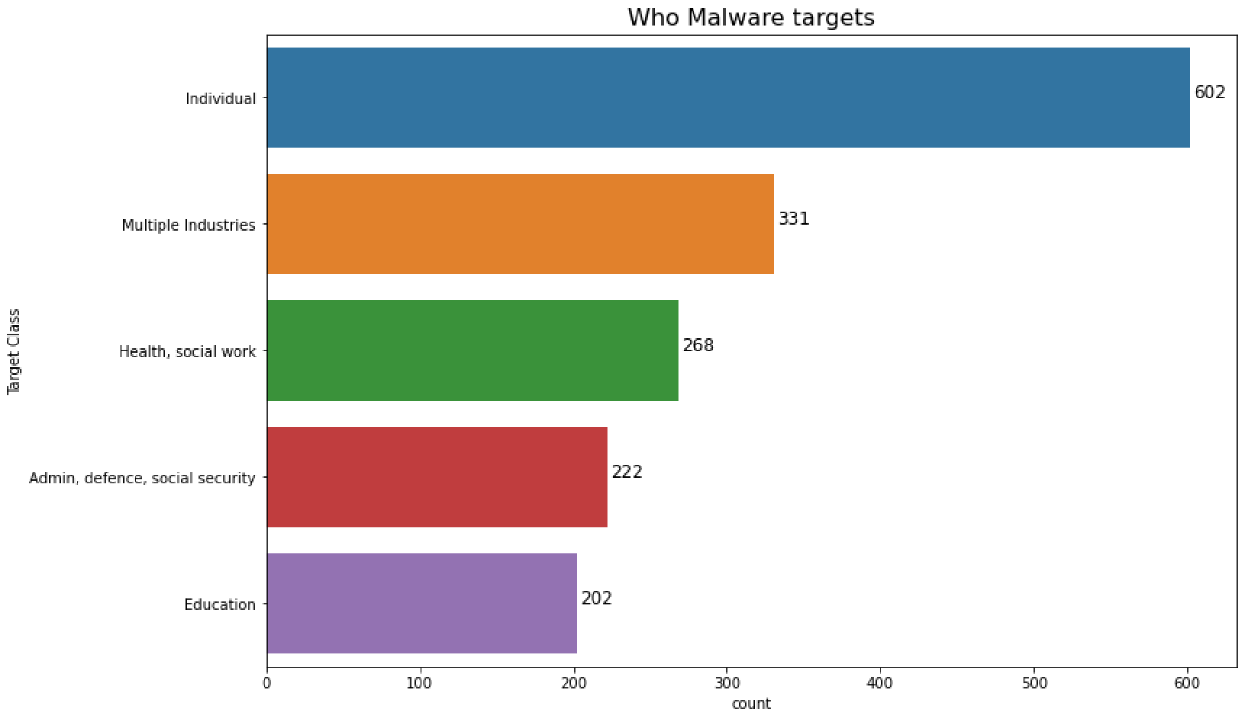


Figure 22: Malware targets

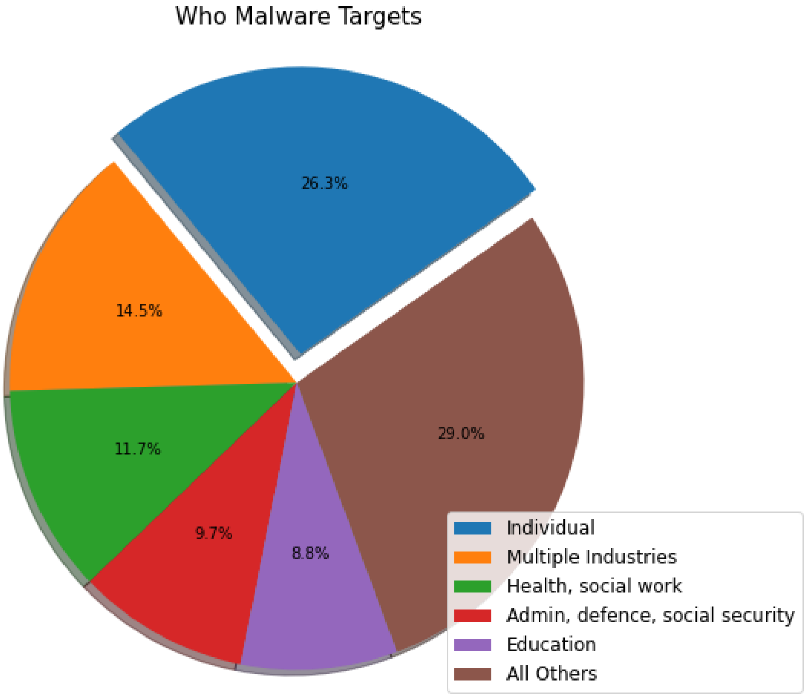


Figure 23: Malware targets percentages

Similar to the attack type analysis, malware also targets individuals the most and industries as the 2nd most frequent. As ‘individuals’ can exist in many kinds, we can go deeper to looking out for what exact individuals are targeted.

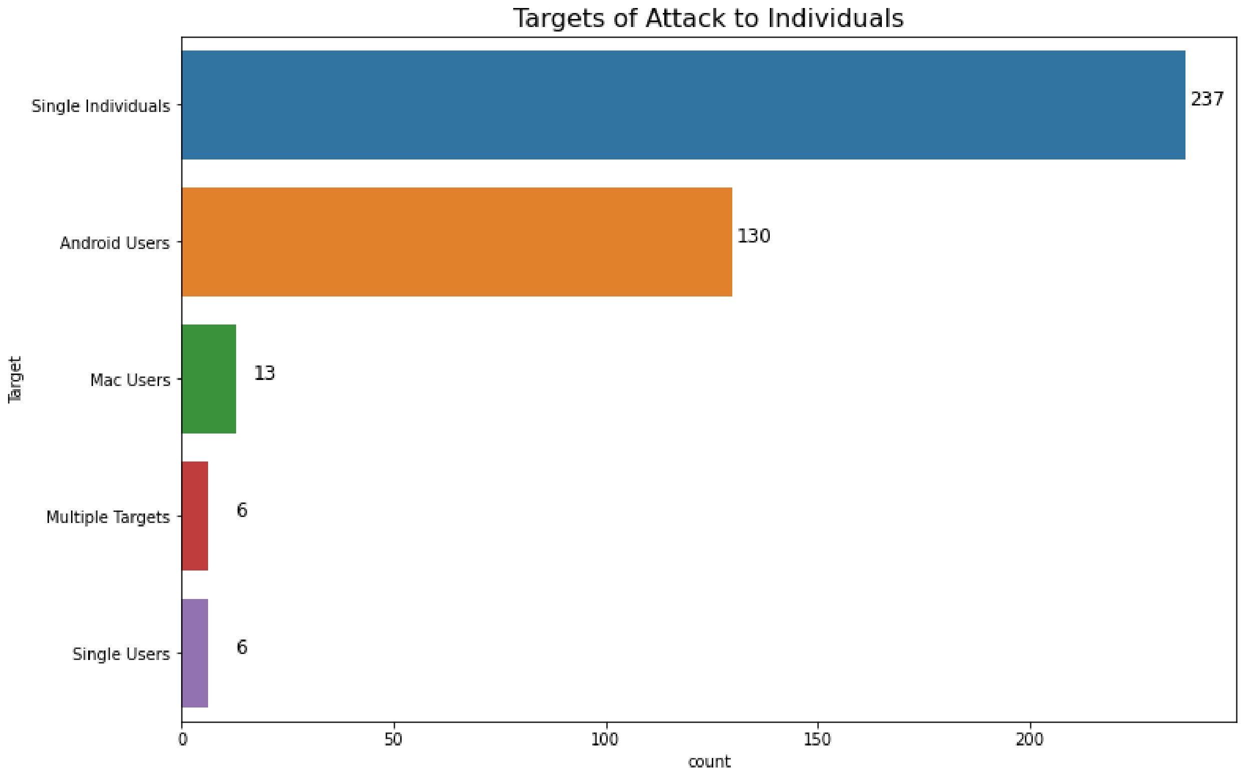


Figure 24: : Types of individuals

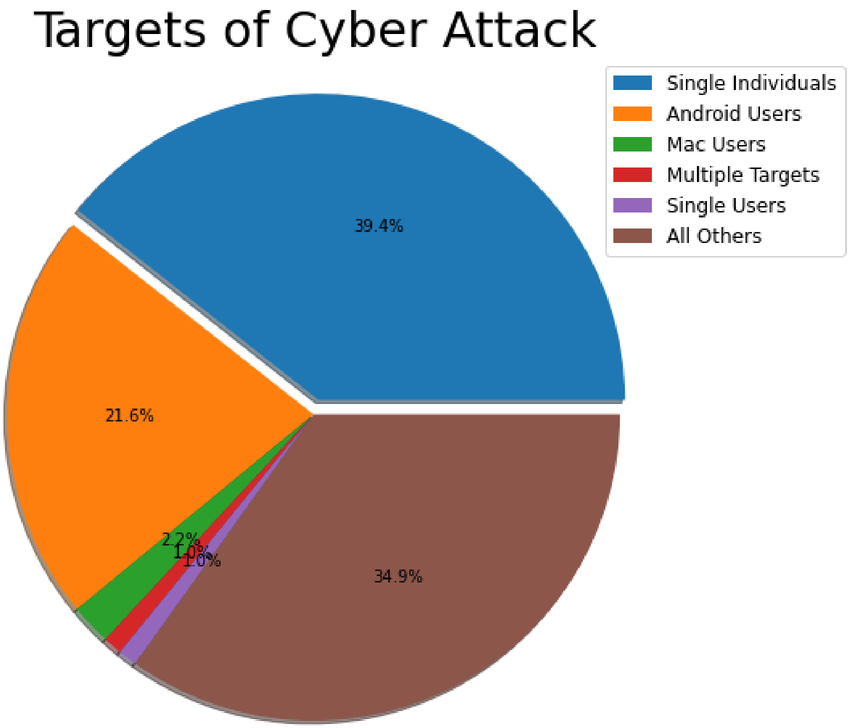
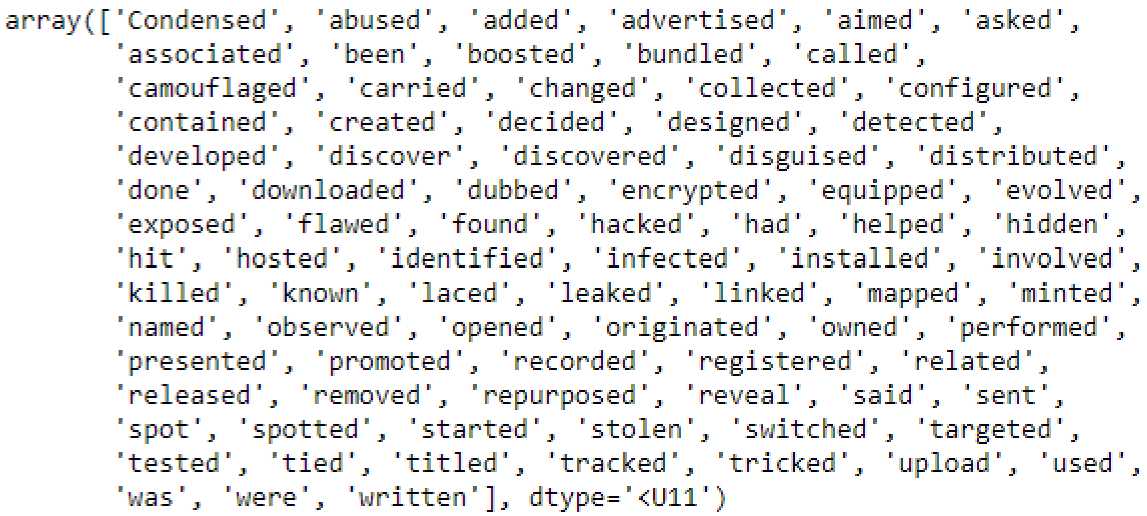


Figure 25: Types of individuals percentages

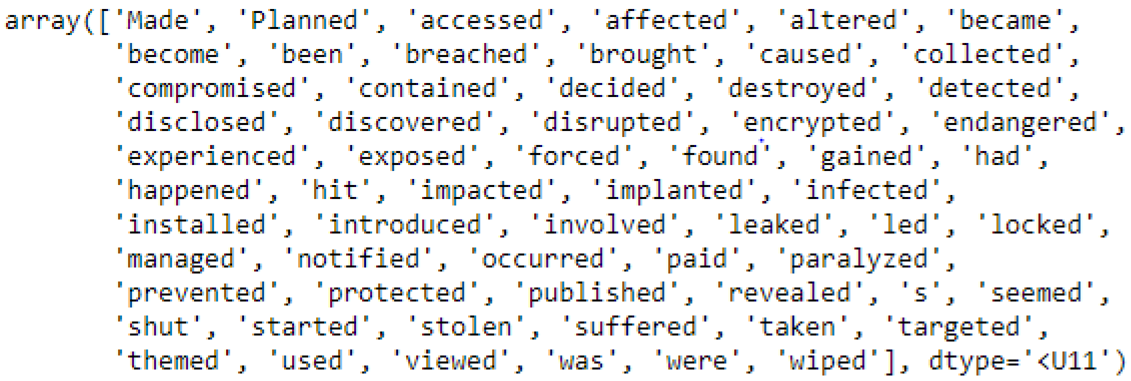
Single individuals and android users are the most frequent targets by malware. Possible reasons may be cyber theft or to cause operation disruption to someone out of hatred or revenge. As for android users, malware could possibly be designed to work easily under android architecture or there may be a wide variety of information sources of malware hacking through android.

1.2. Nltk Part-of-Speech Tagging



```
array(['Condensed', 'abused', 'added', 'advertised', 'aimed', 'asked',
       'associated', 'been', 'boosted', 'bundled', 'called',
       'camouflaged', 'carried', 'changed', 'collected', 'configured',
       'contained', 'created', 'decided', 'designed', 'detected',
       'developed', 'discover', 'discovered', 'disguised', 'distributed',
       'done', 'downloaded', 'dubbed', 'encrypted', 'equipped', 'evolved',
       'exposed', 'flawed', 'found', 'hacked', 'had', 'helped', 'hidden',
       'hit', 'hosted', 'identified', 'infected', 'installed', 'involved',
       'killed', 'known', 'laced', 'leaked', 'linked', 'mapped', 'minted',
       'named', 'observed', 'opened', 'originated', 'owned', 'performed',
       'presented', 'promoted', 'recorded', 'registered', 'related',
       'released', 'removed', 'repurposed', 'reveal', 'said', 'sent',
       'spot', 'spotted', 'started', 'stolen', 'switched', 'targeted',
       'tested', 'tied', 'titled', 'tracked', 'tricked', 'upload', 'used',
       'was', 'were', 'written'], dtype='<U11')
```

Figure 26: Verbs extracted from description column



```
array(['Made', 'Planned', 'accessed', 'affected', 'altered', 'became',
       'become', 'been', 'breached', 'brought', 'caused', 'collected',
       'compromised', 'contained', 'decided', 'destroyed', 'detected',
       'disclosed', 'discovered', 'disrupted', 'encrypted', 'endangered',
       'experienced', 'exposed', 'forced', 'found', 'gained', 'had',
       'happened', 'hit', 'impacted', 'implanted', 'infected',
       'installed', 'introduced', 'involved', 'leaked', 'led', 'locked',
       'managed', 'notified', 'occurred', 'paid', 'paralyzed',
       'prevented', 'protected', 'published', 'revealed', 's', 'seemed',
       'shut', 'started', 'stolen', 'suffered', 'taken', 'targeted',
       'themed', 'used', 'viewed', 'was', 'were', 'wiped'], dtype='<U11')
```

Figure 27: Verbs extracted from description column under USA

From the list of verbs extracted, we can see some of the actions taken in cyber context include 'infected', 'installed', 'disguised', 'encrypted', 'hacked', 'compromised', 'disrupted', 'paralyzed' and ‘stolen'. We can infer that the attacks include methods such as gaining unauthorised access, camouflaging a malicious act into a legitimate one, stealing and locking up data, installing persistensies on the machine and compromising operations.

1.3. Analysing USA

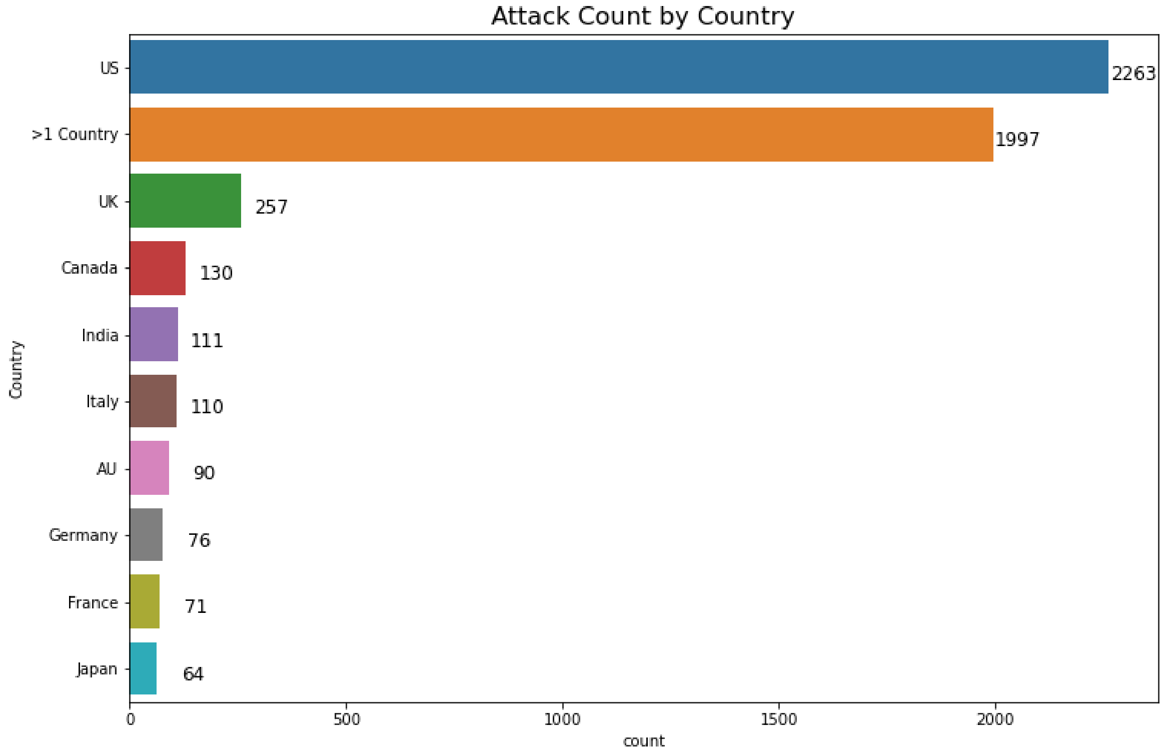


Figure 28: Country attack count

The USA is the country where most attacks are happening. This may be due to the USA being ranked as a “top tier” cyber power, with cyber technology and attacks development originating from them [13].

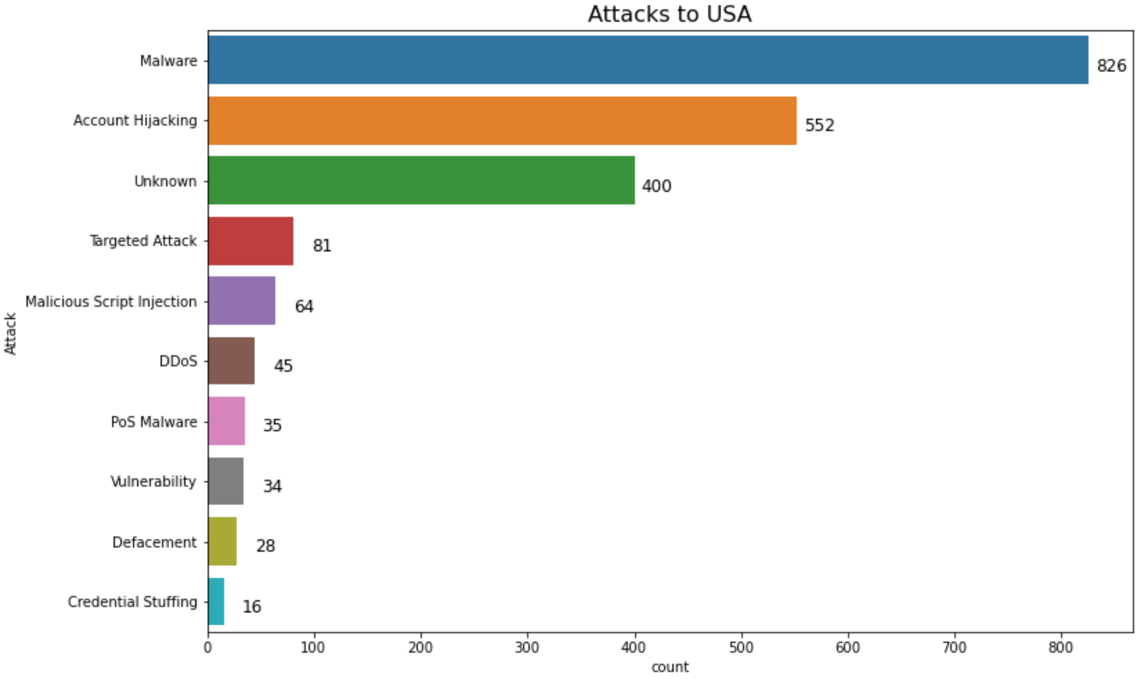


Figure 29: Attack types to USA

Similar to previous observations, malware is also the most frequent attack in the USA.

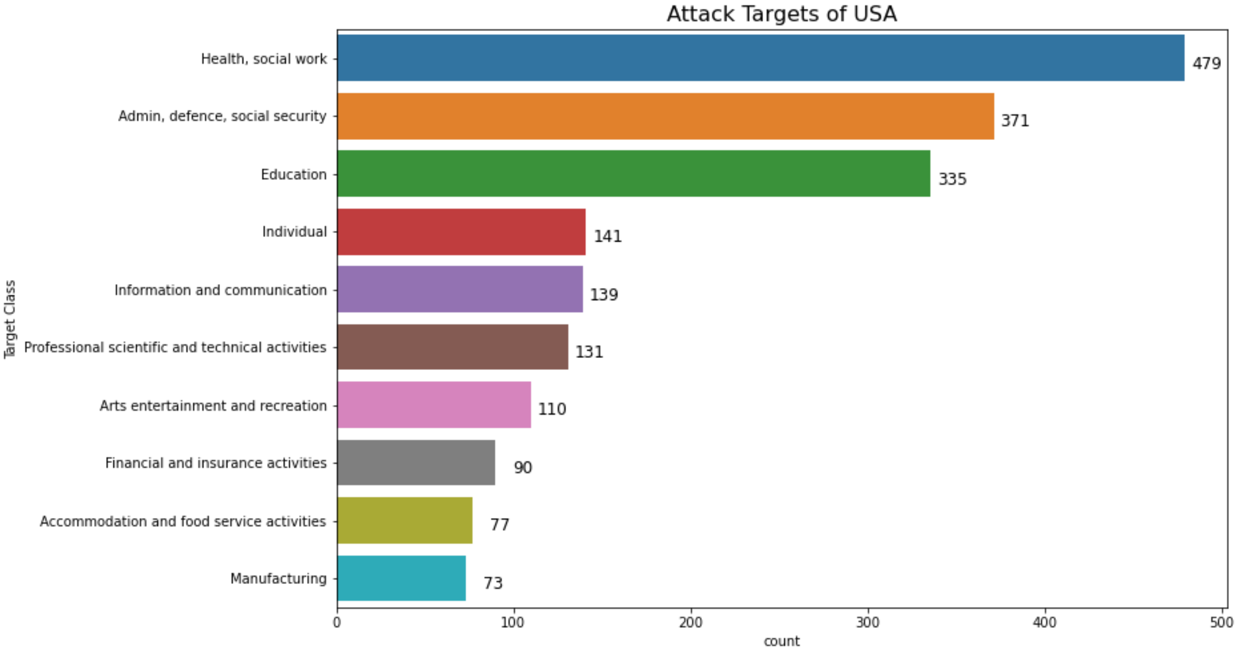


Figure 30: Targets of attack of USA

In the USA, the significant targets of cyber attacks are on health, admin, defence, social and education. It could be due to these firms making more money over there.

2. Veris dataset:

2.1. Trends

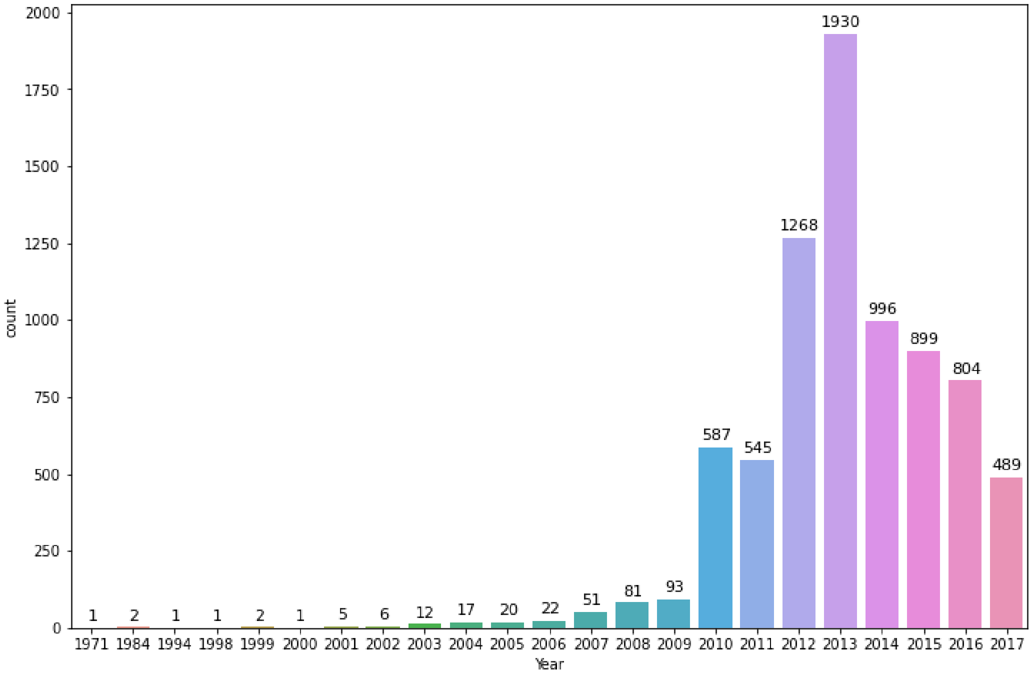


Figure 31: Attack count of veris dataset

Most of the attacks happen between 2010-2017 with 2013 at its peak.

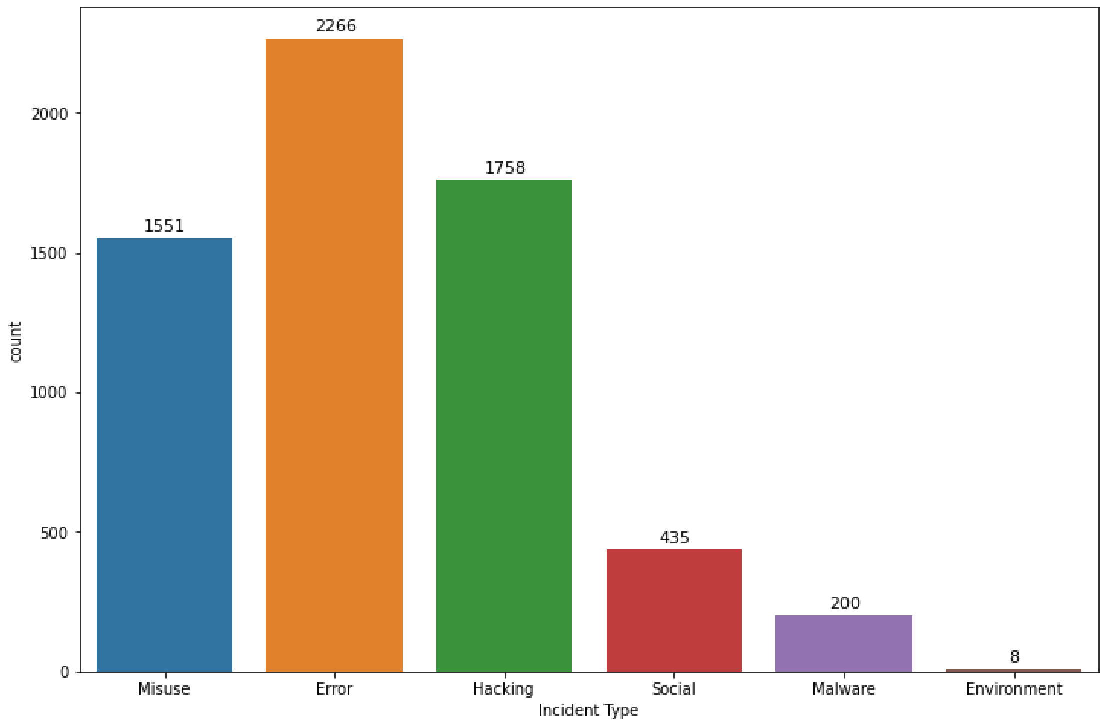


Figure 32: Types of incident

Most of the attacks come under error, misuse and hacking. It could be due during the incident peak period such as in 2013, a large-scale incident such as a major software flaw that led to errors and exploits.

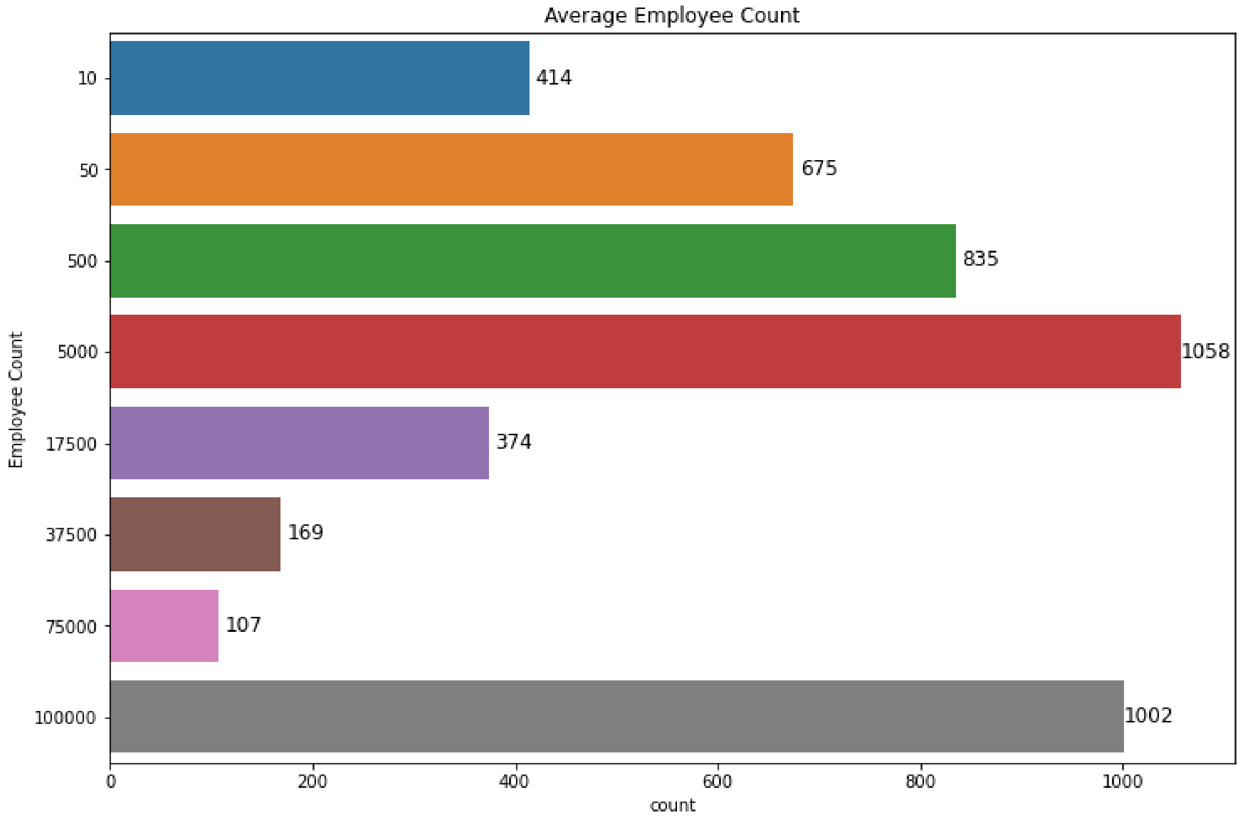


Figure 33: Avg employee count

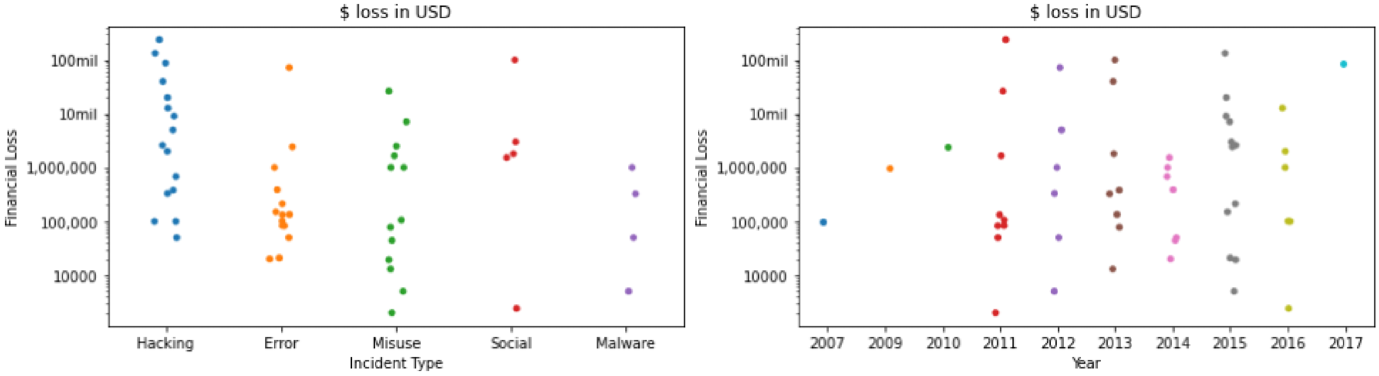


Figure 34: Incident type and year to financial loss

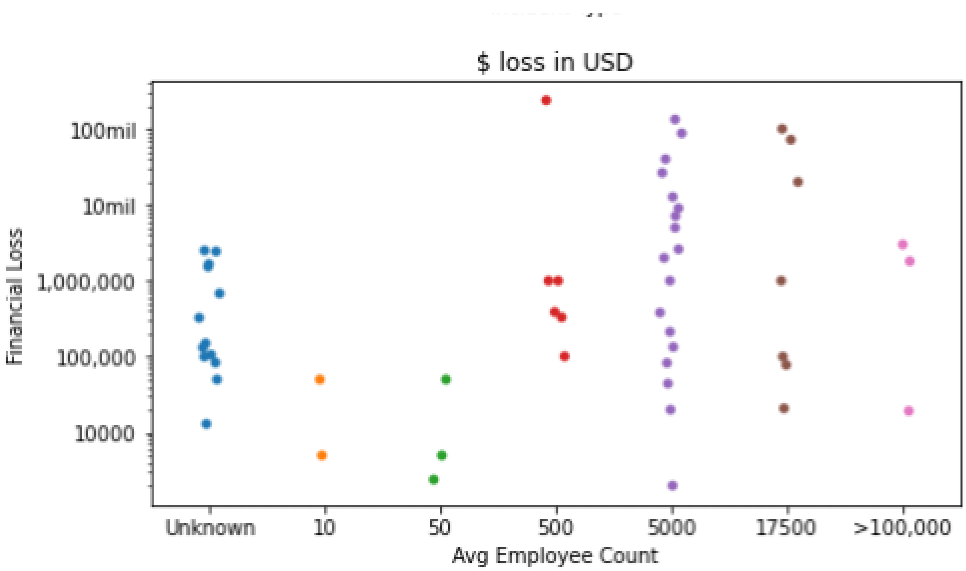


Figure 35: Avg emp count to financial loss

The distributions of the above various categorical factors to financial loss seem random and do not show an inferable pattern or relationship.

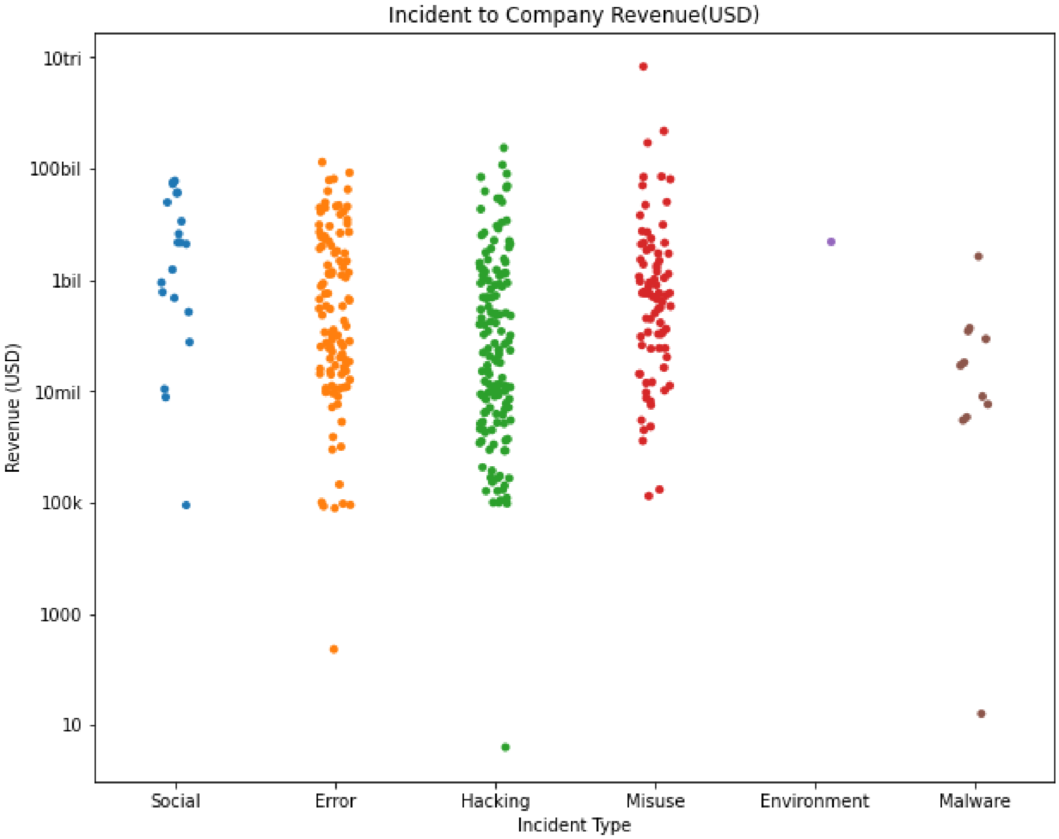


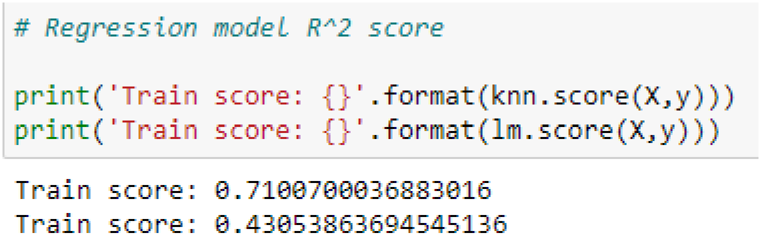
Figure 36: Incident type to company revenue

Most of the company's revenue lies between 100k to 100billion, with a few outlier exceptions.

2.2. Machine Learning

2.2.1. Predicting Losses

Using linear regression and k-nearest neighbor regression, the financial loss is predicted based on the incident type, employee count and year of incident. The metric used to measure the accuracy of the model is the R-squared. R-squared is a statistical measure of how close the data are fitted to the regression model, which is the proportion of the variance for a dependent variable that is explained by an independent variable. The higher the value, the better the model is. The best possible score is 1.0 and can be negative which means the model performs worse. The accuracy on R squared error is shown below without splitting the dataset to train/test sets.

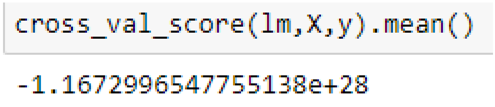


```
# Regression model R^2 score

print('Train score: {}'.format(knn.score(X,y)))
print('Train score: {}'.format(lm.score(X,y)))
```

```
Train score: 0.7100700036883016
Train score: 0.43053863694545136
```

Figure 37: R^2 of trained model

Although the scores of both models are positive, further validation shows that the model is overfitting as test sets of K nearest neighbor model and cross validation of the linear regression model shows negative accuracy. This means that the models will not work accurately on new ‘unseen’ data.



```
cross_val_score(lm,X,y).mean()
```

```
-1.1672996547755138e+28
```

Figure 38: Negative r^2 on cross validation

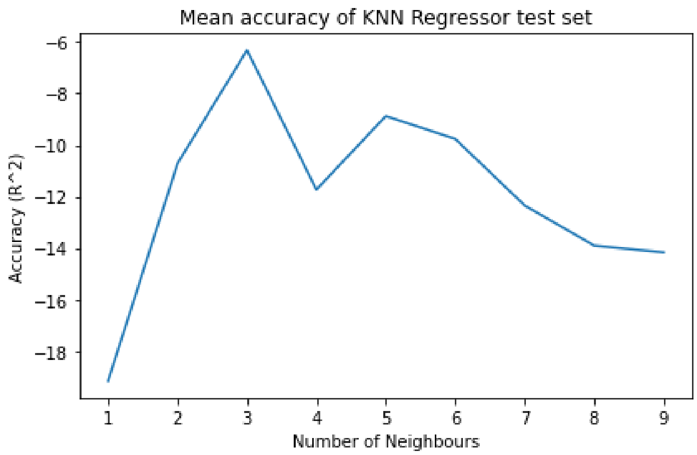
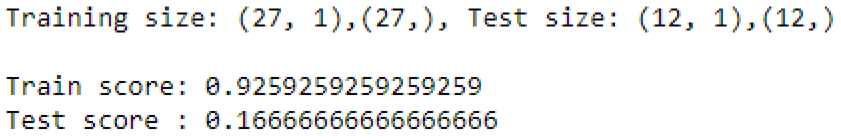


Figure 39: Negative r^2 for test set

Overfitting is a situation when the data relies too much on the trained model and cannot be used to predict new data. This is expected as there are only 39 data points. A possible future remedy is to obtain more data so that the weightage of accuracy for each data is significantly reduced.

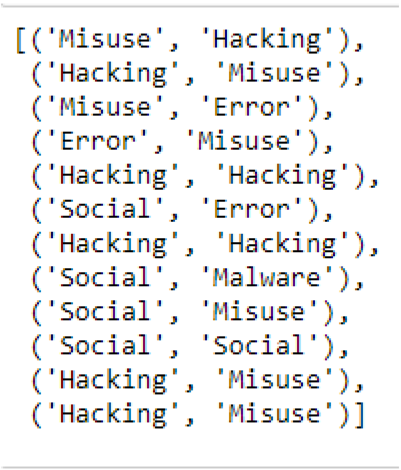
2.2.2. Classifying incident type

A decision tree model is used to predict what an incident type will be given the amount of losses a company suffered.



```
Training size: (27, 1),(27,), Test size: (12, 1),(12,)

Train score: 0.9259259259259259
Test score : 0.16666666666666666
```

Figure 40: Data size and accuracy of classifier



```
[('Misuse', 'Hacking'),
 ('Hacking', 'Misuse'),
 ('Misuse', 'Error'),
 ('Error', 'Misuse'),
 ('Hacking', 'Hacking'),
 ('Social', 'Error'),
 ('Hacking', 'Hacking'),
 ('Social', 'Malware'),
 ('Social', 'Misuse'),
 ('Social', 'Social'),
 ('Hacking', 'Misuse'),
 ('Hacking', 'Misuse')]
```

Figure 41: Predicted(left) vs actual(right)

Due to the small amount of data, there is overfitting and lack of accuracy on the test set. Out of the 12 test samples, only 5 of them were predicted correctly.

3. Github dataset:

3.1. Trends

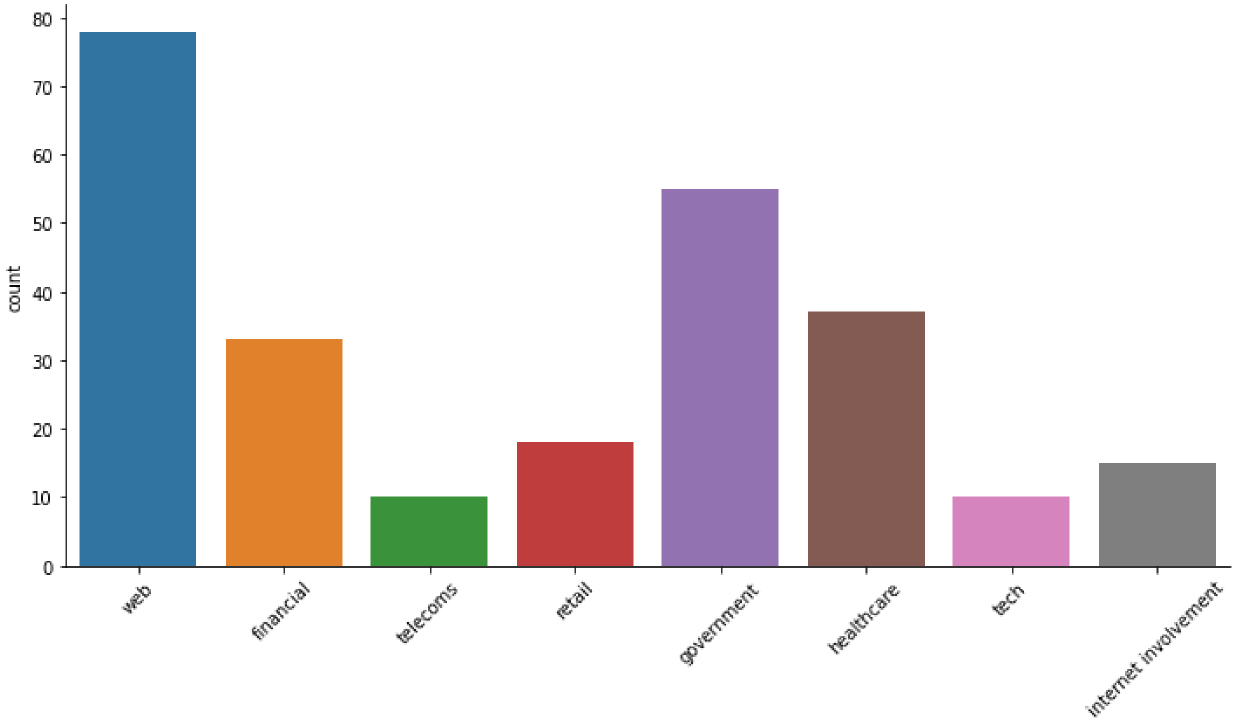


Figure 42: Type of organisation

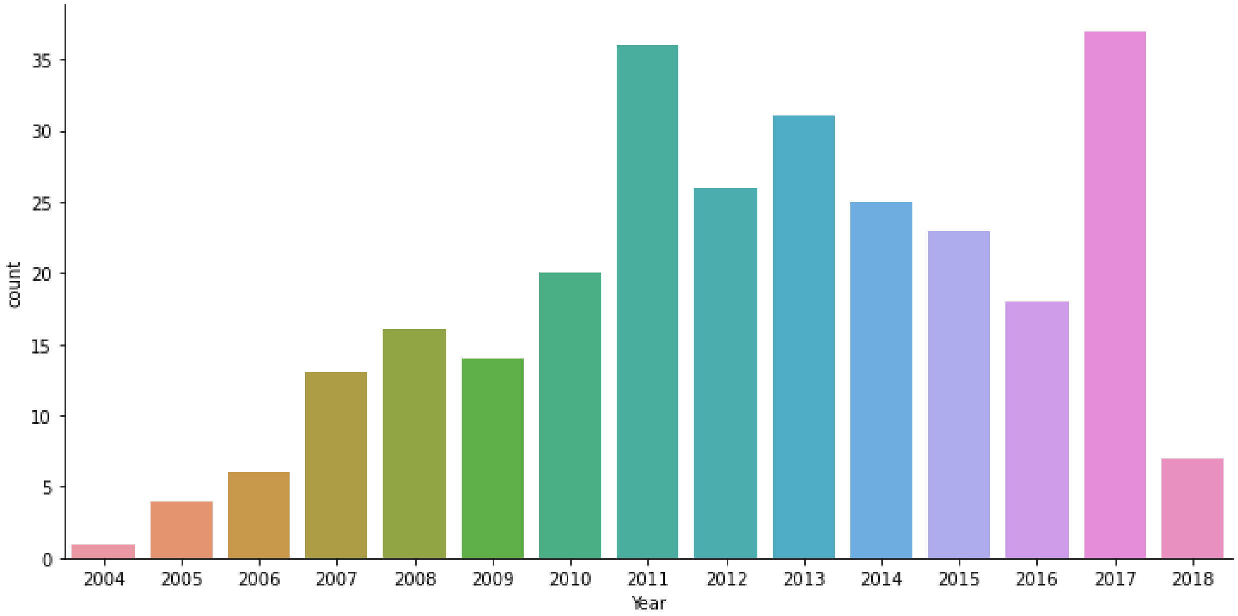


Figure 43: Year of breach

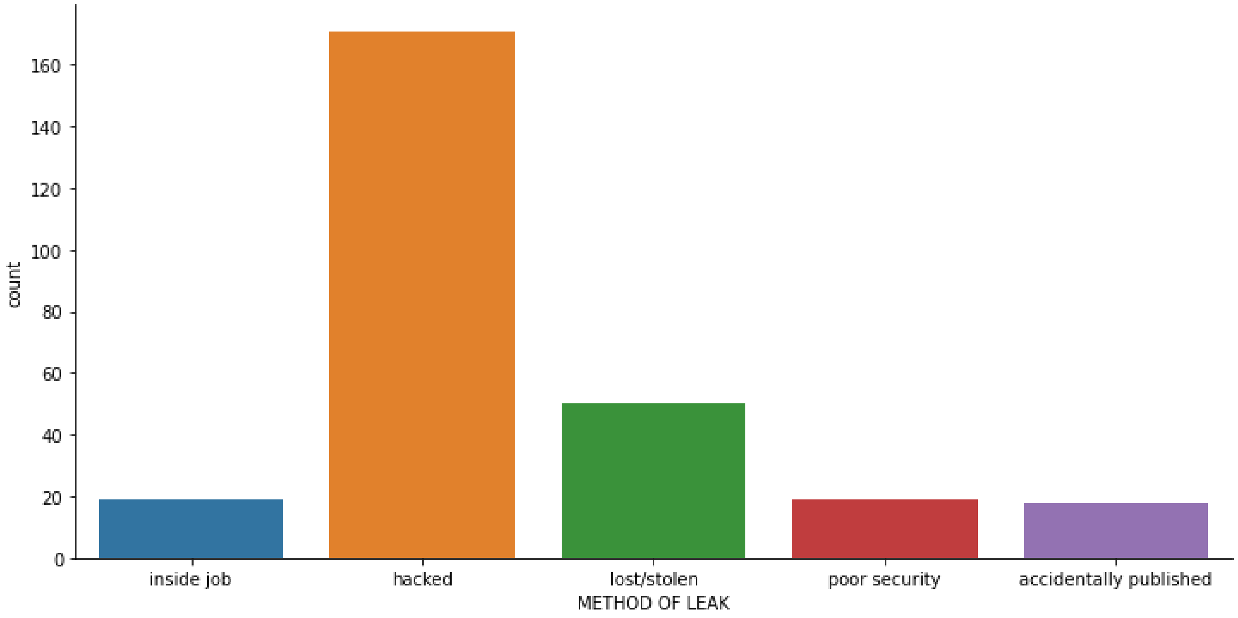


Figure 44: Method of data leak

Above shows the distributions of the type of organisations present in the data breach, year of breach and method of data leak. Hacking is the most frequent type and is most likely the cause of web attacks.

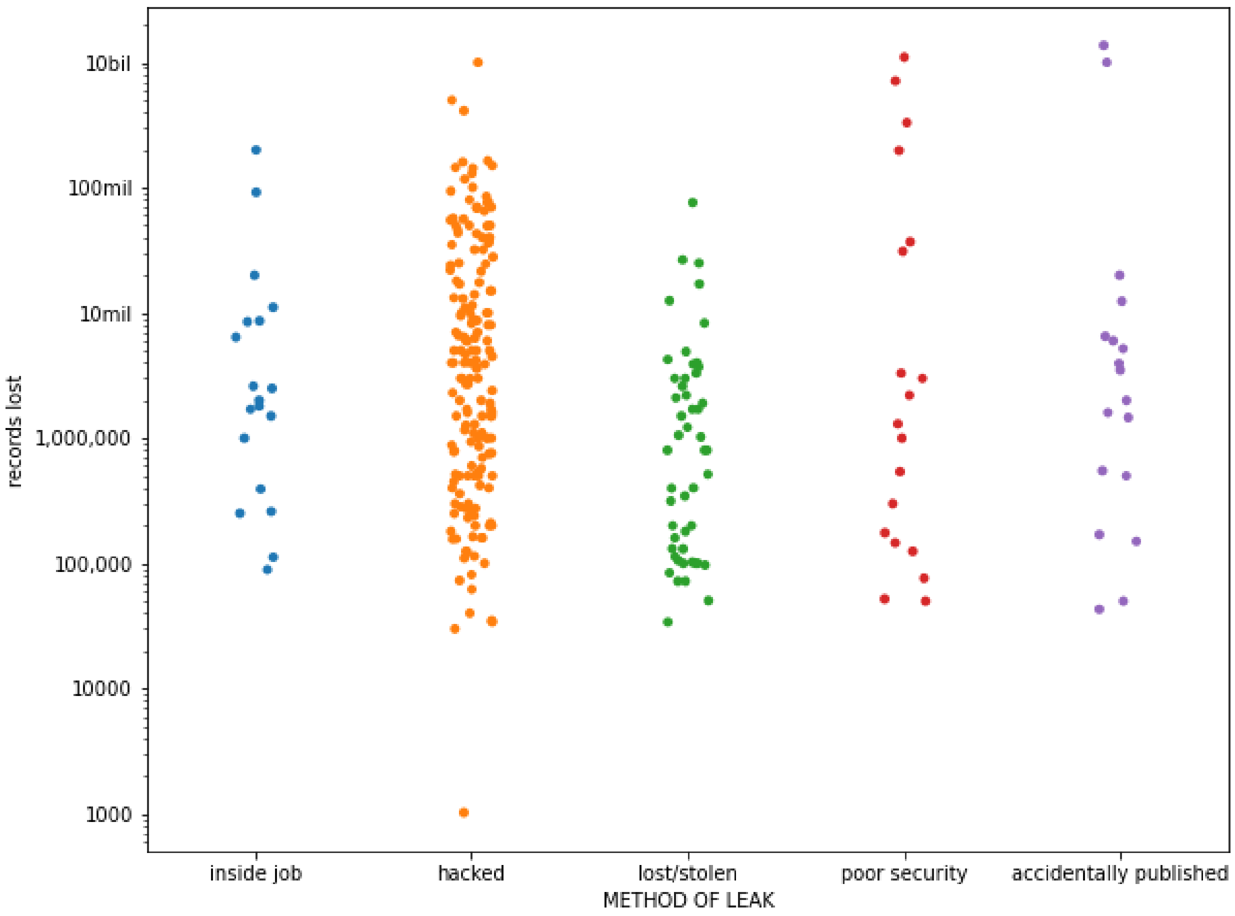


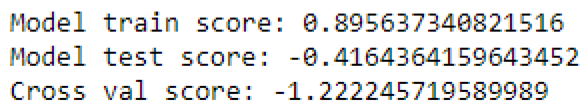
Figure 45: Method leak to records lost

Many of the records lost come under hacking as it may be the most effective form of attack.

3.2. Machine Learning

3.2.1. Predicting records lost

The records lost are predicted based on method of leak, year, organisation and data sensitivity using the K nearest neighbors regression model.



```
Model train score: 0.895637340821516
Model test score: -0.4164364159643452
Cross val score: -1.222245719589989
```

Figure 46: R^2 score

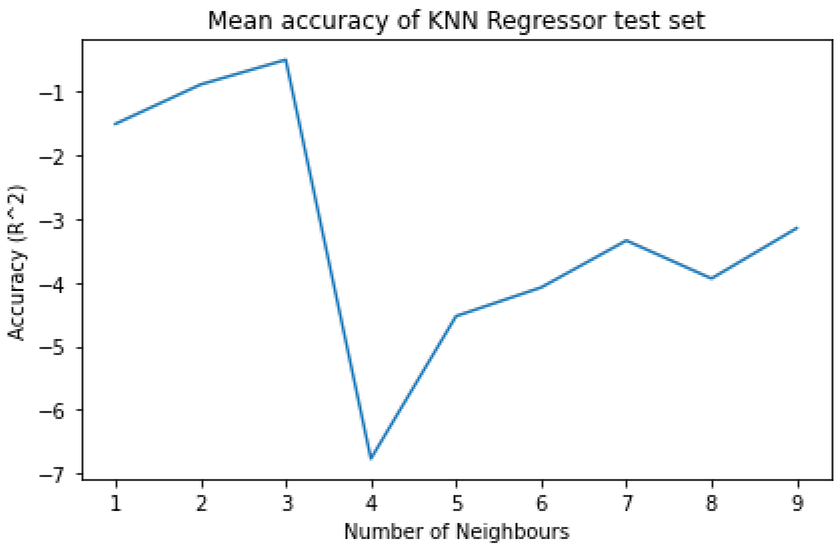
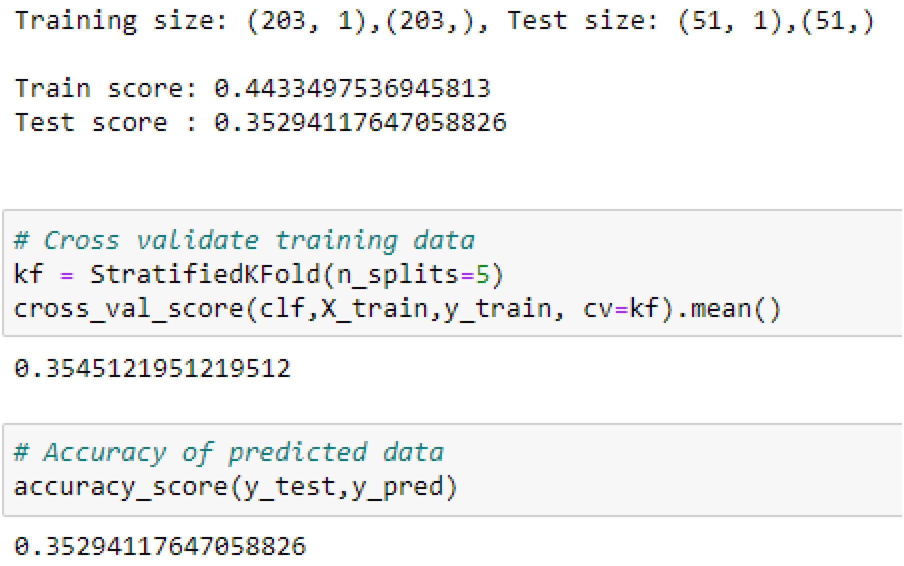


Figure 47: Accuracy of different no. neighbors

Due to the lack of data, the model faced the problem of overfitting as the accuracy score is negative for the test set and cross validation. The accuracy is also negative when testing with different amounts of K neighbors.

3.2.2. Classifying type of organisation

A decision tree classifier is built to predict the type of organisation to be hit by a data breach given the year of attack is the predictor.



```
Training size: (203, 1),(203,), Test size: (51, 1),(51,)

Train score: 0.4433497536945813
Test score : 0.35294117647058826
```

```
# Cross validate training data
kf = StratifiedKFold(n_splits=5)
cross_val_score(clf,X_train,y_train, cv=kf).mean()
```

```
0.3545121951219512
```

```
# Accuracy of predicted data
accuracy_score(y_test,y_pred)
```

```
0.35294117647058826
```

Figure 48: Accuracy of classifier model

The model turns out to be performing not very well as predicted values are only 35% accurate. This can be further improved if there is more data.

Conclusion

The objective of this project is to visualise cyber incidents trends from various data sources and to perform predictive modelling as accurately as possible. We can learn from analysing the above 3 data sources that attack occurrences are generally on the rise over the past decade. Common attacks are often malware and hacking targeted towards organisations and people with the USA being one of the most frequent countries experiencing cyber attacks. Since the USA is the current superpower globally, they are the ones that will be developing new techniques themselves to address their own cyber threats.

Limitations

The main issue of this project was the lack of data for each individual dataset, especially when predictive modelling needs to be done. The absence of statistics hinders the exploration of machine learning analysis which provides very insightful and useful information. Many important factors have missing values such as revenue and financial loss in the veris dataset. Filtering out inconsistencies in variables used for predictive modelling, there are only 39 data points left which could potentially lead to underfitting/overfitting issues. Most of the attributes over the other 2 datasets are categorical and descriptive data, which only can be visualised through distributions and NLP tools. There are also very limited openly available cyber incident data sources, some have little information that would not be fruitful for analysis. There are also many online incident news that only provide description and lack statistical information.

Another issue could the credibility of the data. There are doubts regarding the validity / biasness of the data from the author. One example is that most incidents in the hackmageddon dataset are from the USA. This may be due to the author living in the USA and only curates data relevant to his/her country and does not check in detail about other countries. Another example may be the author having more data but did not openly include them in the dataset to falsely show an untrue trend to achieve any kind of agenda. The information shown from analysis cannot be fully trusted as many other cyber attack trends online show different results on a specific topic.

Future work

It would be useful if there could be an implementation of a centralised repository meant for analysing cyber incidents, more statistics rather than just descriptive data. Data analysts that work with cyber security can also be more open to security incidents and provide more data sources and keep the public educated on the importance of modern day cyber security.

References

[1]
Johnson, J., 2021. Cyber crime: reported damage to the IC3 2020 | Statista. [online] Statista. Available at: <https://www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us/> [Accessed 12 September 2021].

[2]
Paganini, P., 2021. How Threat Hunters Operate in Modern Security Environments. [online] Security Affairs. Available at: <https://securityaffairs.co/wordpress/73274/security/cyber-security-threat-hunters.html> [Accessed 12 September 2021].

[3]
HACKMAGEDDON. 2021. *Home*. [online] Available at: <https://www.hackmageddon.com/> [Accessed 12 September 2021].

[4]
Veriscommunity.net. 2021. *the veris community database (vcdb)*. [online] Available at: <http://veriscommunity.net/vcdb.html> [Accessed 12 September 2021].

[5]
GitHub. 2021. *awghatevaibhavi - Overview*. [online] Available at: <https://github.com/awghatevaibhavi/> [Accessed 12 September 2021].

[6]
Tutorialspoint.com. 2021. *Matplotlib - Bar Plot*. [online] Available at: <https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm> [Accessed 9 October 2021].

[7]
Tutorialspoint.com. 2021. *Matplotlib - Bar Plot*. [online] Available at: <https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm> [Accessed 9 October 2021].

[8]
Seaborn.pydata.org. 2021. *seaborn.countplot — seaborn 0.11.2 documentation*. [online] Available at: <https://seaborn.pydata.org/generated/seaborn.countplot.html> [Accessed 9 October 2021].

[9]
Seaborn.pydata.org. 2021. *seaborn.scatterplot — seaborn 0.11.2 documentation*. [online] Available at: <https://seaborn.pydata.org/generated/seaborn.scatterplot.html> [Accessed 9 October 2021].

[10]
Analytics Vidhya. 2021. *K-Nearest Neighbors Algorithm | KNN Regression Python*. [online] Available at: <https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/> [Accessed 9 October 2021].

[11]
Analytics Vidhya. 2021. *Decision Tree Classification | Guide to Decision Tree Classification*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/> [Accessed 9 October 2021].

[12]
Nltk.org. 2021. *5. Categorizing and Tagging Words*. [online] Available at: <https://www.nltk.org/book/ch05.html> [Accessed 9 October 2021].

[13]
India Today. 2021. India a third-tier country in cyber warfare capabilities, report says US more powerful than China. [online] Available at: <https://www.indiatoday.in/technology/news/story/india-a-third-tier-country-in-cyber-warfare-capabilities-report-says-us-more-powerful-than-china-1820261-2021-06-28> [Accessed 12 September 2021].