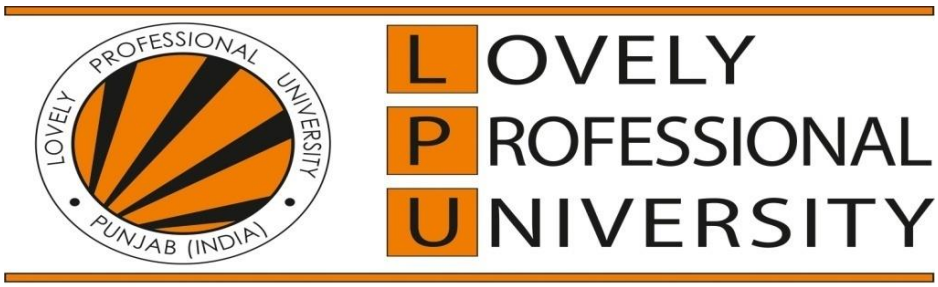


DATA MANAGEMENT PROJECT REPORT

ON

Analysis of Cyber Attacks Happened All Over the World

Submitted by:

Manas Bhardwaj

11704118

Under the guidance of:

Sandeep Kaur - 23614

School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Manas Bhardwaj** bearing Registration number **11704118** has completed **Data Management (INT217)** project titled, **"Analysis of Cyber Attacks Happened All Over the World"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
School of Computer Science and Engineering
Lovely Professional University
Date:

DECLARATION

I, **Manas Bhardwaj**, student of **Computer Science and Engineering** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature:

Registration No. 11704118 Manas Bhardwaj

ACKNOWLEDGEMENT

I take this opportunity to present our votes of thanks to all those guideposts who really acted as lightening pillars to enlighten my way throughout this project that has led to the successful and satisfactory completion of this Project. I am grateful to **Lovely Professional University** for providing us with an opportunity to undertake this project and providing us with all the facilities. I am highly thankful to All for their active support, valuable time and advice, whole-hearted guidance, sincere cooperation and painstaking involvement during the project and in completing the assignment of preparing the said project within the time stipulated. Lastly, I am thankful to all those, particularly the various friends, who have been instrumental in creating a proper, healthy and conducive environment and including new and fresh innovative ideas for me during the project, without their help, it would have been extremely difficult for me to complete the project in a time-bound framework.

INDEX

INTRODUCTION

1. Data:

It is the raw and isolated facts about an entity or records.

2. Information:

When the data are used to extract some insights then those data which are being used is called informative data and the results that come out of them is called information.

3. Data Analysis:

The process of analysing the data which later turns into information is called data analysis. In this process we observe the data to make some inference out of it, we use several tools and resources to read the jargon datasets into the common and usable form.

Or, It is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names while being used in different business, science, and social science domains.

4. Data Science:

The science of studying the behaviours of the data like what they stand for, what can be extracted from them, their nature and how they can be used for future references and used to solve some real-world problems. It is one of the emerging fields of science and technologies where gazillions of data are fetched into a machine and analysed for the beautiful insights.

5. Excel 2016:

Microsoft office tools which are used widely for data analysis and exploration. Excel 2016 is the best tool for data analysis when it comes to handy data sets. Excels contains all the features for converting small uncleaned data into a readable form and also contains different functions which provide an eminent role in data exploration. It's easy to use and cosmopolitan in the distribution.

6. Tableau Prep:

Tableau Prep is another exciting tool which is widely used by data analysts around the world. It is used when we have to deals with a lot of inbuilt errors, like removing wrong spellings, removing the redundancy of the data, removing data using different aggregate functions and visual guidance. Sometimes we have to deal with hidden errors and there this tool comes in handy experiences. Tableau is not only used to explore the errors but also used to explore and analyse the data sets for better insights.

SCOPE OF ANALYSIS

Data visualization is useful in a plethora of areas, whether it be technical sector or not technical sector. Every micro-business or macro large scale industries require this analysis to meet their different standards and goals. Data visualization is not only required in one field but in the today world, it has become quintessential in all most all the sectors.
Data visualization is used in business, managerial skills, for training and testing machine models, to solve real-world problems with the aid of machine learning and artificial intelligence. Moreover, it has been able to solve some of the complex problems of the world.

Some directions to take when exploring the data:

1. Top 5 Cyber-attacks happened on 5th October 2019.
2. Top 5 attacking countries.
3. Attack type analysis.
4. Countries Attacking Too & Fro.
5. Top 5 victim countries.
6. The number of attacks happened on the different hour of the day.
7. Type of Attacks happened on the different hour of the day.

Aim of this project is to answer the above objectives in the form of visualization by creating a dashboard to convey the answers effectively.

ETL PROCESS

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL full-form is Extract, Transform and Load.

Extraction:
Extraction is the first step in the ETL process, which basically deals with the data mining and data selection phase. In this step, we have to use different sources through which we are going to extract our files for the analysis. Extraction involves the files like .csv and others which need cleaning. This process plays eminent roles because one we are done with the extraction, everything turns out to be simple for us. Different functions are used to maintain the normality and atomicity of the data sets.

Transformation:
Transformation is the second phase of data cleaning. Once the data are extracted and cleaned it needs to be transformed into a set of cleaned and well-defined data. The transformation takes place in different models. In our coursework, we did transformation with an in-built Pivot Table and manually too. However, Pivot table helped a lot in the data transformation. Different charts models are used to make the visual appearance of the data sets.

Load/Visualization:
Data once get transformed needs to be analysed. This process is the ultimate process in which results are extracted from the graphs and visual models. Different models represented dynamically can be used to anticipate results, make careful observations and create a fascinating story out of it.

DATASET DESCRIPTION:

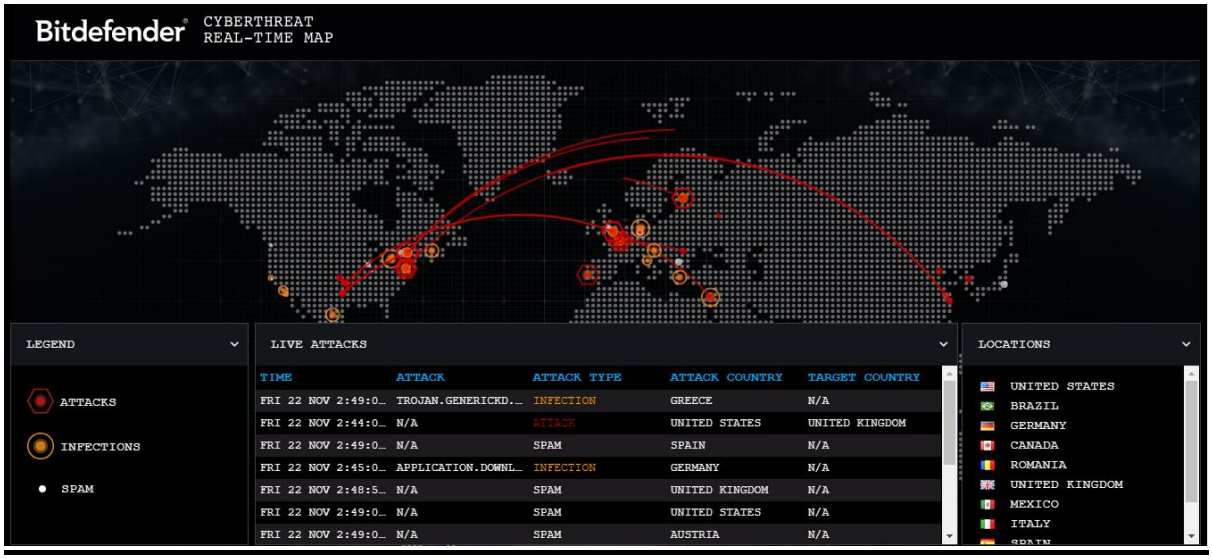
Source :	https://threatmap.bitdefender.com/
Date of Collection :	05-Oct-19
Time of Collection :	00 : 00 - 23 : 00
Collected By :	Manas Bhardwaj
Number of Columns :	6
Number of Rows :	22,447
Method of Collection :	Web Scraping using python, refer Scraping-script.py in root directory
Tools used for cleaning :	Tableau prep, Python, Microsoft Excel 2016
Method Used to remove NULLS :	Exclusion of rows,Forward Filling in pandas(Python).

Code Book

Columns Name	Description	Type	Total NULLS
Time	It contains time of the attack in 24-hour format.	Time(HH) 24-hour	0
Date	It contains date of attack.	Date(DD-mm-YYYY)	0
Attack Name	It contains the name of attack which was performed on the victim country by the attacking country.	Text	0
Attacking Country	It contains the name of the attacking country.	Text	0
Target Country	It contains the name of the target country on which attack was performed.	Text	0
Attack Type	It contains the type of attack, i.e INFECTION, SPAM, ATTACK.	Text	0

DATA SOURCE:

Link: https://threatmap.bitdefender.com



Data Collection Script:

```
# This script is written to scrape the data from a dynamic website.
# A dynamic website is a site that contains dynamic pages such as templates,
# contents, scripts etc. In a nutshell, the dynamic website displays various
# content types every time it is browsed. The web page can be changed with the
# reader that opens the page, character of consumer interplay, or day time.
# Or, Dynamic website is a website in which the data changes very frequently.
```

```
from bs4 import BeautifulSoup
from selenium import webdriver
from time import sleep
import csv
```

```
# Below code is to automate Chrome only if you want to automate Mozilla you have
# to download geckodriver according to the version you are using of Mozilla.
# Just download the geckodriver and place it in the directory
# and replace 'chromedriver.exe' with the geckodriver in below line.
# Or, you can simply uncomment the line 12 and comment the line 11.
```

```
driver = webdriver.Chrome(executable_path='chromedriver.exe')
driver.get('https://threatmap.bitdefender.com/') # Opens the link in the browser.
sleep(60) # It gives the website sufficiant time to load all its content.
while(1):
```

Executes the javascript to get the HTML every time it is executed without
refreshing the website since we are scraping the dynamic website it is
very important to prevent refreshing of the website since the data will not
be available as we wanted.

```
    res = driver.execute_script("return document.documentElement.outerHTML")
    # To convert the response of the above line to readable html.
    soup = BeautifulSoup(res,'lxml')
    # It find the table-body('tbody') tag since we want the data from the table
    # which is updating regularly and the id of the table is 'attacks_data'.
    # In futue it can change since website is maintained regularly so you can
    # change the below body tag and id tag accordingly.
    tags = soup.find_all('tbody',{'id':'attacks_data'})
    tags = tags[0] # We just want first table.

    for tr in tags.find_all('tr'): # Iterate over the all the rows in the table.
        print(tr.contents,end='\n') # Prints the content of the row in form of list.

        # To Collect the data we use 'csv' module to write in csv file.
        with open("attacks.csv",'a+',newline='') as f:
            writer = csv.writer(f,delimiter=',',quotechar='|', quoting=csv.QUOTE_MINIMAL)
            if len(list(tr.contents)) == 5: # we want that data only which has all the 5 columns data.
                writer.writerow(list(tr.contents)) # Writes the row to the csv file.

    sleep(20)
driver.quit() # When you're done just press cntrl+c to quit and the browser will get close.
```

Uncleaned Data Collected in CSV format,

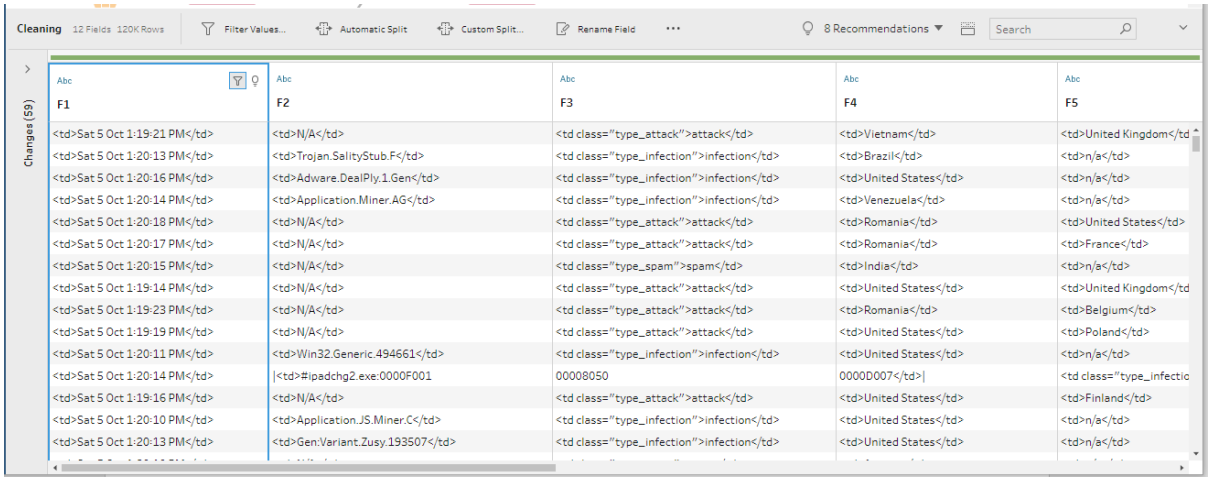
```
[<td>Fri 22 Nov 2:34:14 AM</td>, <td>N/A</td>, <td class="type_attack">attack</td>, <td>Romania</td>, <td>Finland</td
[<td>Fri 22 Nov 2:34:16 AM</td>, <td>#esif_assist_64.exe:0000F001,00008006,00008050,00009903</td>, <td class="type_in
[<td>Thu 21 Nov 7:44:31 PM</td>, <td>Trojan.Downloader.VBS.HY</td>, <td class="type_infection">infection</td>, <td>Br
[<td>Fri 22 Nov 2:34:41 AM</td>, <td>N/A</td>, <td class="type_spam">spam</td>, <td>Costa Rica</td>, <td>n/a</td>]
[<td>Fri 22 Nov 2:33:45 AM</td>, <td>N/A</td>, <td class="type_attack">attack</td>, <td>Vietnam</td>, <td>United Stat
[<td>Fri 22 Nov 2:33:45 AM</td>, <td>N/A</td>, <td class="type_attack">attack</td>, <td>France</td>, <td>Germany</td>
[<td>Fri 22 Nov 2:34:39 AM</td>, <td>Application.BitCoinMiner.SX</td>, <td class="type_infection">infection</td>, <td
[<td>Fri 22 Nov 2:34:41 AM</td>, <td>N/A</td>, <td class="type_spam">spam</td>, <td>United States</td>, <td>n/a</td>]
```

DATA CLEANING:

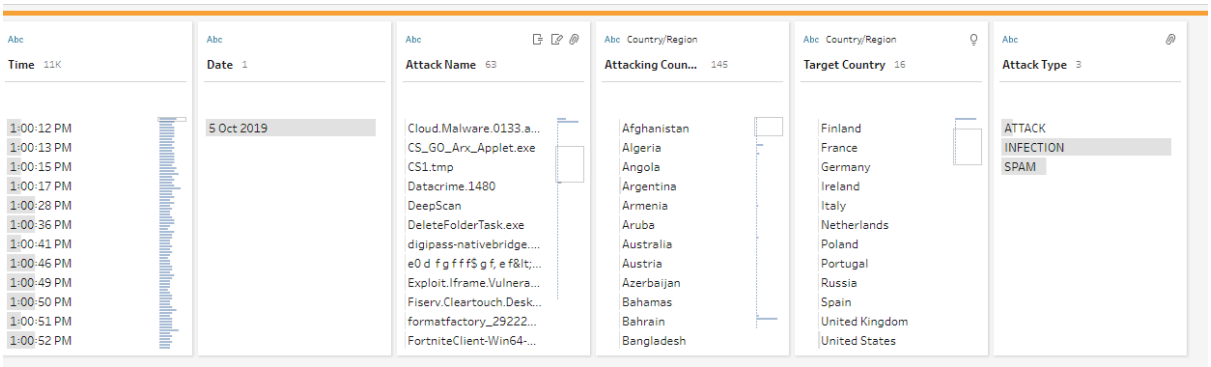
For cleaning the messy data tableau prep was used,

Fig. Tableau Flow File

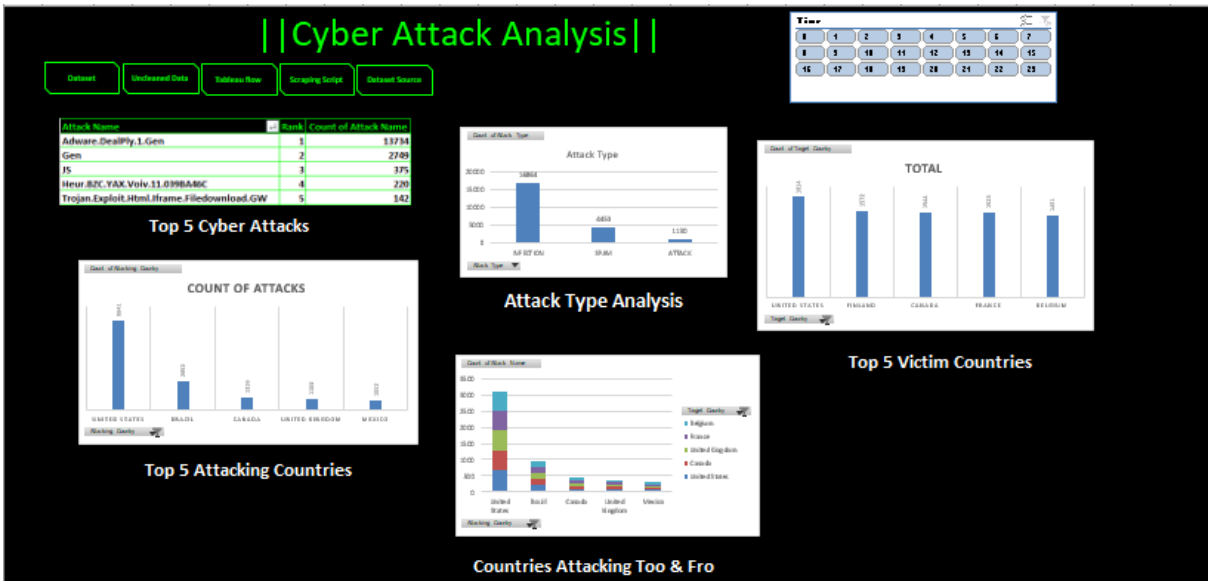
Messy Data:



Cleaned Data:



DASHBOARD



ANALYSIS OF DATASET

1. **Top 5 Cyber-attacks happened on 5th October 2019.**

In the analysis, it was found that the following were the attacks that were performed in 24-hours.

Adware was the most used to attack the victims' device, you go online with your nice, well-behaved browser, only to see it fly into a virtual tantrum, as an onslaught of advertisements either pops up, slides in from the side, or otherwise inserts itself to interrupt and even redirect your intended activity. And no matter how much you click to close those windows, they keep buzzing you like flies at a picnic.

Since we all visit a lot of websites it's very easy for the attacker to infect through adware.

To protect yourself from adware Use caution and practice safe computing. That means thinking twice before immediately downloading and installing any new software—especially freeware.

Attack Name	Rank	Count of Attack Name
Adware.DealPly.1.Gen	1	13734
Gen	2	2749
JS	3	375
Heur.BZC.YAX.Voiv.11.039BA46C	4	220
Trojan.Exploit.Html.Iframe.Filedownload.GW	5	142

Fig. Pivot table showing top 5 attacks

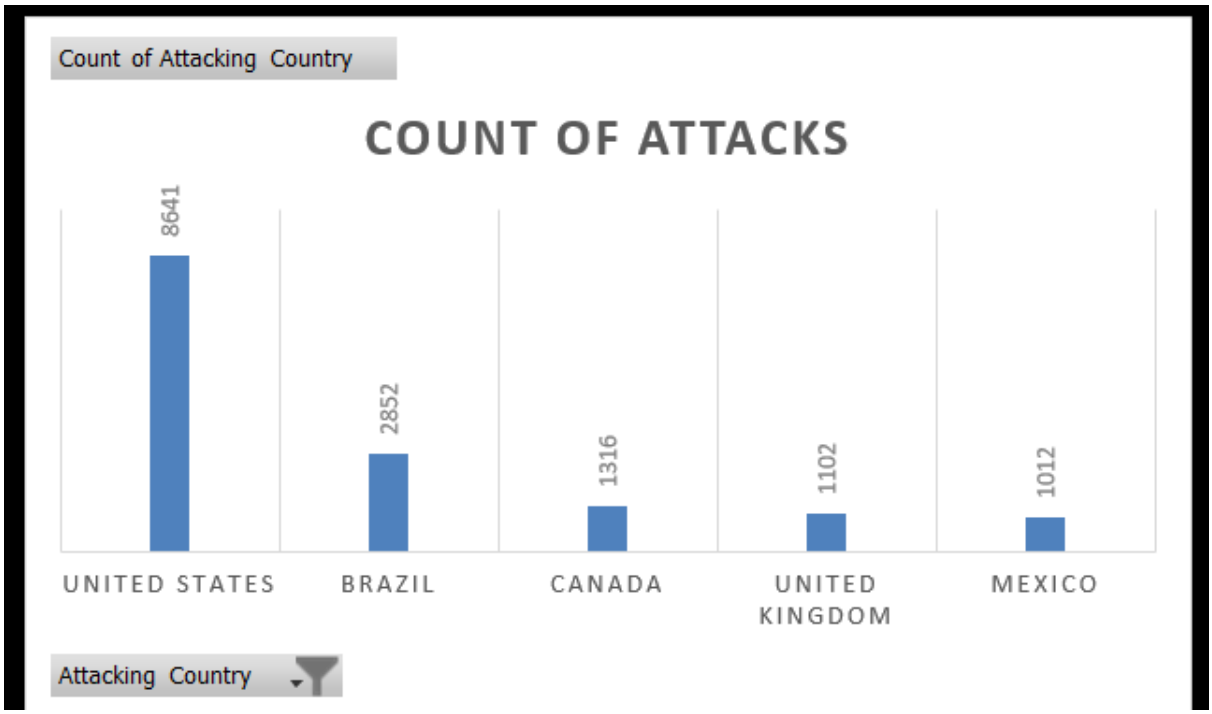
2. **Top 5 attacking countries.**

United states being the most developed country attacked most number of time to different countries in 24 hours. United States is highly dependent on the Internet and therefore greatly exposed to cyber-attacks. At the same time, the United States has substantial capabilities in both defense and power projection thanks to comparatively advanced technology and a large military budget.

According to the, CNCERT report on U.S. cyber-attacks – nations around the world are increasingly using cyber as both an economic weapon and a military tool to achieve strategic objectives. For smaller nations, cyber-attacks enable an asymmetric response to much larger rivals. And for larger nations, cyber-attacks give military planners and politicians another tool in the toolbox to exert influence, all without being forced to respond with military means.

Attacking Countries	Count of Attacking Country
United States	8641
Brazil	2852
Canada	1316
United Kingdom	1102
Mexico	1012
Grand Total	14923

Fig. Pivot table showing the countries who attacked the most.

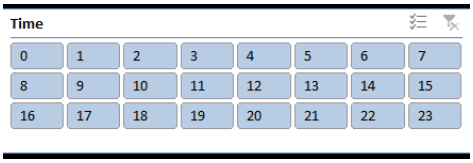


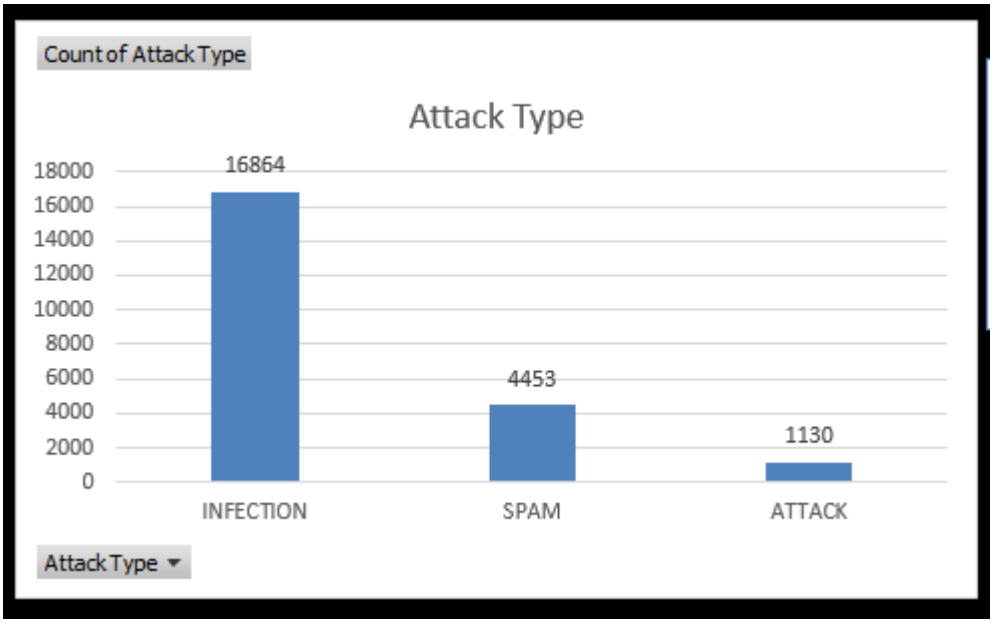
Visualization

3. Attack type analysis.

Since we all know that internet is booming at a very fast pace, we visit a lot of websites in a day. We download a lot of files which can easily be injected with payload and victim can be easily exploited.

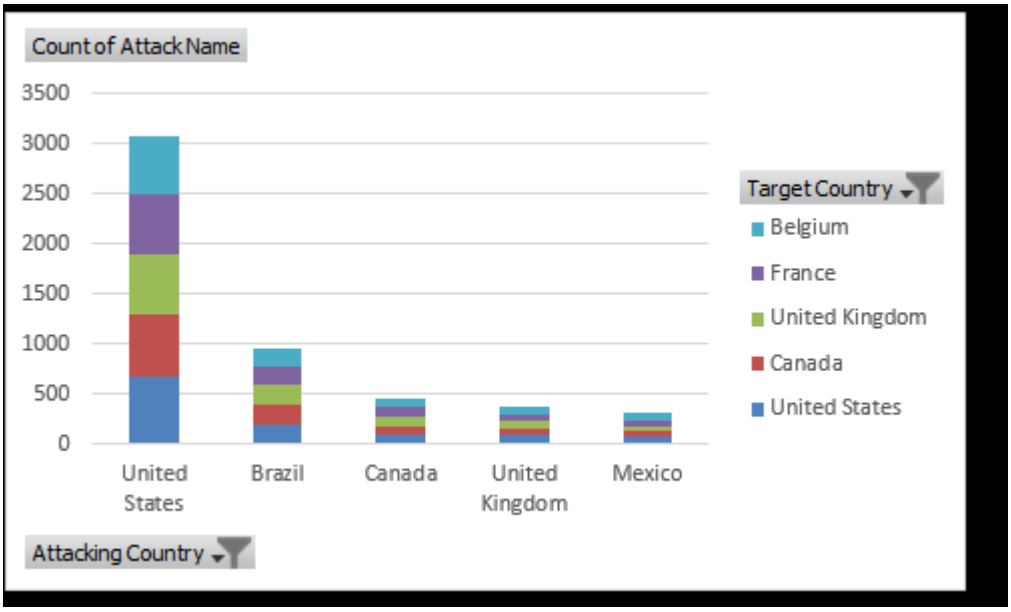
Attack Type	Count of Attack Type
INFECTION	16864
SPAM	4453
ATTACK	1130
Grand Total	22447





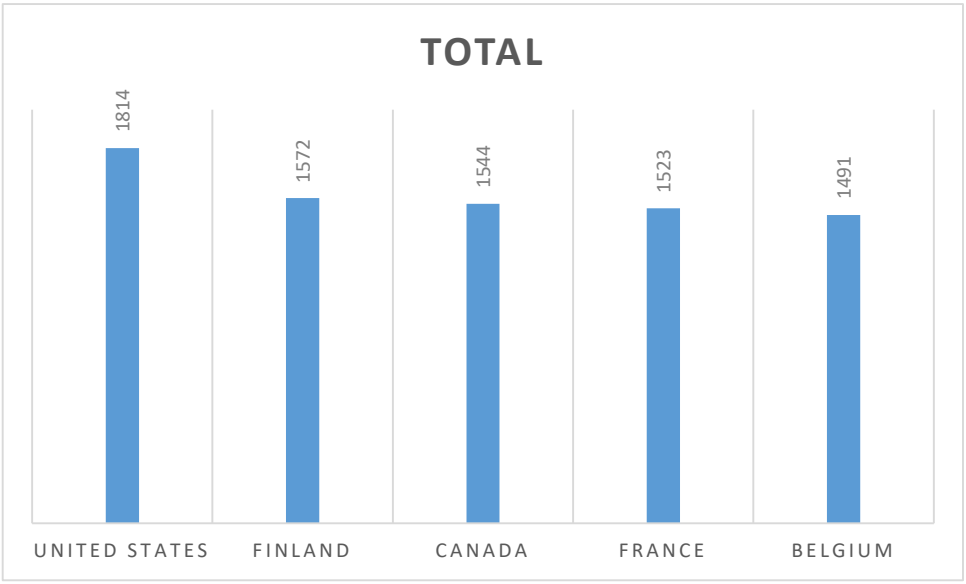
4. Countries Attacking Too & Fro.

Count of Attack Name	Target Countries					
Attacking Countries	United States	Canada	United Kingdom	France	Belgium	Grand Total
United States	670	616	608	604	582	3080
Brazil	198	199	192	180	189	958
Canada	86	94	95	92	77	444
United Kingdom	91	69	66	75	70	371
Mexico	66	57	57	55	84	319
Grand Total	1111	1035	1018	1006	1002	5172



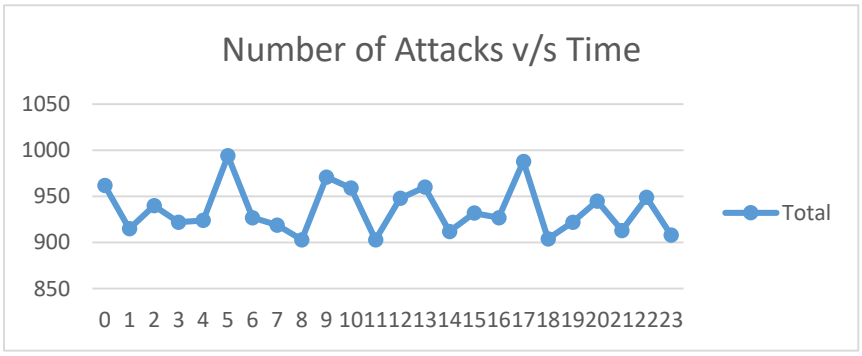
5. **Top 5 victim countries.**

Row Labels	Count of Target Country
United States	1814
Finland	1572
Canada	1544
France	1523
Belgium	1491
Grand Total	7944



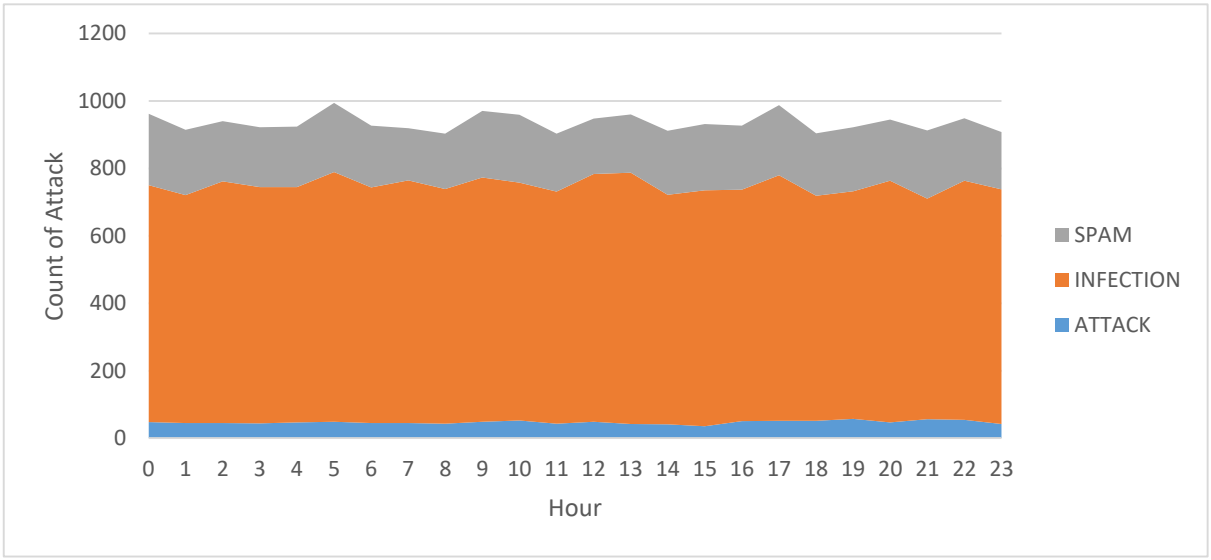
6. **The number of attacks happened on the different hour of the day.**

Hour	Count of Attac
0	962
1	915
2	940
3	922
4	924
5	994
6	927
7	919
8	903
9	971
10	959
11	903
12	948
13	960



7. **Type of Attacks happened on the different hour of the day.**

Count of Attack Name	Attack Types			
Hours	ATTACK	INFECTION	SPAM	Grand Total
0	48	702	212	962
1	45	676	194	915
2	45	717	178	940
3	44	701	177	922
4	47	698	179	924
5	49	740	205	994
6	45	699	183	927
7	45	720	154	919
8	43	696	164	903
9	49	724	198	971
10	52	706	201	959
11	43	688	172	903



References and Bibliography

1. https://selenium-python.readthedocs.io/
2. https://help.tableau.com/current/prep/en-us/prep_about.htm
3. https://www.tableau.com/learn/tutorials/on-demand/getting-started-tableau-prep
4. https://www.youtube.com/watch?v=e_o2S3oJnYQ
5. https://www.tableau.com/learn/tutorials/on-demand/tableau-prep-interface
6. https://www.youtube.com/watch?v=YDvpgdy2ox8
7. https://www.youtube.com/watch?v=rwbho0CgEAE
8. https://www.youtube.com/watch?v=RdTozKPY_OQ
9. https://www.excel-easy.com/
10. https://www.tutorialspoint.com/excel/index.htm
11. https://www.guru99.com/excel-tutorials.html