National College of Ireland

Data Warehousing and Business Intelligence Project

on

Analysis of Cyber threats by Proportion of Online Population

Shikhar Srivastava
x18106960

MSc/PGDip Data Analytics – 2018/9

Submitted to: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Shikhar Srivastava
Student ID:	x18106960
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Simon Caton
Submission Due Date:	26/11/2018
Project Title:	Analysis of Cyber threats by Proportion of Online Population

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	November 26, 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Analysis of Cyber threats by Proportion of Online Population

Shikhar Srivastava
x18106960

November 26, 2018

Abstract

The world is moving rapidly towards a digital age and with it, the number of people having access to internet is also growing day by day. This in turn means that a lot of IP addresses are present in the world today, and with this, several problems associated with the internet like attacks and hacking are also increasing. The project tries to identify maximum possible number of malicious IP addresses in Europe, Latin America, Asia Pacific and Africa and analyze if there is any relationship between these IP addresses and the number of internet users in respective countries. Project also tries to identify if the countries of Europe, Latin America, Asia Pacific and Africa are present in the top countries list, categorized by 7 categories – Lowest Rate of Malware Infection, Highest Rate of Malware Infection, Highest Number of Users Attacked by Ransomware, Highest Percentage of DDoS Attacks, Least Amount of Freedom on Net, Highest Amount of Freedom on Net, Most Vulnerable Countries to Hacking. The above goal has been achieved by gathering data from different sources and consolidating them in a star schema using R, integrating them in SSIS, SSAS and reporting in Microsoft Excel.

1 Introduction

Internet today is very important and an integral part of our life. While the internet helps us in many ways, there are several malicious activities which happen on internet nowadays that can affect us in multiple ways and the rate of these activities is increasing day by day. It has become very important specially now, to protect our data. Almost all companies today are taking cyber security very seriously because of recent major malicious activities like ransomware attacks. These activities are not just limited to giant firms, but also target common people. IP address plays a very important role in everything which is related to internet, or, one can say a very crucial role. Many people have contributed a lot to make the internet a secure place, yet malicious activities tend to happen again and again. There already are few approaches which can efficiently detect malicious websites[1], or, cluster IP addresses based on their interaction with other IP addresses[2]. One of the studies state that proactive detection of malicious IP packets can even predict DDoS attacks[3], and there are many more. Above studies prove that the internet is not very safe and also contribute towards the motivation of this project. Project based on malicious IP addresses has gathered data from 4 different sources and try to show the AS-IS condition while trying to identify relationship between malicious

Source	Type	Brief Summary
Abuseipdb.com	Semi-Structured	Malicious IPs are extracted from this data source.
Internetworldstats.com	Structured	Country wise internet users are present here.
Comparitech.com	Unstructured	Top countries based on 7 factors extracted from an image present here
Statista	Structured	LATAM countries by internet users

Table 2: Summary of sources of data used in the project

IPs and percentage of internet users by country. It has gathered data based on top countries by different parameters like 'Countries with Highest Number of Users Attacked with Ransomware' etc.

(Req-1) Relationship of Country wise Internet Users with Number of IP reports in Europe.

(Req-2) Relationship of Countries having "Highest Number of Ransomware Attacks" with IP reports.

(Req-3) Relationship of Categories under which IPs have been recently reported to Countries with 'Highest Rate of Malware Infection'.

2 Data Sources

Project uses data from 4 different data sources which is then transformed according to project needs and consolidated in a star schema in SSAS. First source is abuseipdb.com from where malicious IPs alon with details like ISP, Domain, Usage Type have been extracted using R code [REFER R CODE]. Categories under which people have recently reported these IPs (as being malicious) are also extracted from this website along with the number of times it has been reported in respective category. Second source of data is internetworldstats.com from where country wise number of internet users have been taken for Europe, Asia and Africa. Data from this source has been manually organised in required structure. Third source is compritech.com and from this source data has been extracted from an image using an online OCR tool `www.onlineocr.net`. This data source provides data for top countries based on these factors - Highest Rate of Malware Infection, Lowest Rate of Malware Infection, Highest Amount of Freedom on Net, Least Amount of Freedom on Net, Highest Number of Users Attacked with Ransomware, Highest Percentage of DDoS attacks, Most Vulnerable Countries to Hacking. Fourth source of data is statista from where LATAM countries by number of internet users have been extracted which has then been merged into data from internetworldstats.com in SSIS workflow.

2.1 Source 1: Abuseipdb.com

Malicious IP dataset was scraped from: `https://www.abuseipdb.com/sitemap?page=1` by iterating the page number in the url through R code[REFER R CODE]. Then data w.r.t each IP was fetched from: `https://www.abuseipdb.com/check/62.210.199.83` by passing all IPs extracted from first url in a for loop in R code.

Fields fetched from above urls are: IP, ISP, Usage Type, Domain Name, Country, City, Reported, FirstReported, Lastreported

However, relevant to this project are: IP, Country, Reported

Another dataset was extracted from the same source using the same R Code[REFER R Code] which has following fields: IP, Categories, n(). n() here is nothing but count of IP reports in respective category.

However, relevant to this project are: Category, n()

Another dataset which was extracted from this source is: `https://www.abuseipdb.com/categories` From here category description is being fetched for each category

Although, Category description is a part of the data model, it is not being used anywhere to answer anything. It was just included to describe a particular category if anybody is having confusion figuring out category by name.

2.2 Source 2: Internetworldstats.com

Country wise internet users data for Europe, Asia and Africa was copied and pasted form here in excel: `https://www.internetworldstats.com/stats4.htm#europe` `https://www.internetworldstats.com/stats3.htm#asia` `https://www.internetworldstats.com/stats1.htm`

Fields fetched from above urls are: Country, Population, Internet Users 31-Dec-2017 All 3 fields are used in the project and this downloaded dataset is latest by 31-Dec-2017

2.3 Source 3: Compritech.com

In this data source, data has been extracted from: `https://cdn.comparitech.com/wp-content/uploads/2017/02/Countries-that-are-the-Most-and-Least-Cyber-Safe-final.jpg`

Data from thisnimage has been extracted using an online OCR tool: `www.onlineocr.net` Data ectracted from this image include fields: Lowest Rate of Malware Infection, Highest Rate of Malware Infection, Highest Number of Users attacked with Ransomware, Highest percentage of DDoS attacks, Least Amount of Freedom on net, Highest Amount of Freedom on net, Most Vulnerable Countries to Hacking

However, relevant to this project are: Highest Rate of Malware Infection, Highest Number of Users attacked with Data extracted from this website was by published on 25-Aug-2018

2.4 Source 4: Statista

Fourth dataset which has been used here is from statista: `https://www.statista.com/statistics/186919/number-of-internet-users-in-latin-american-countries/`

This dataset contains LATAM countries by internet users. Fields used from this dataset are: Country, Internet Users in Million. Both these fields are being used in the project.

However, two blank columns were manually added in this source: Population and Percent Internet Users because other countries were having these 2 columns included in the dataset when data was copied from internetworldstats.com, so to combine all of them in a single table 2 blank columns were added manually in this dataset in excel.

Dataset publishing date is mentioned as to be January 2018

3 Related Work

This project gives an overall view of how on a country level, number of reports on IPs is related to other factors like percent online population. It shows the AS-IS image of malicious activities on a global level. There are several other works, not exactly in malicious IPs, but related to malicious domains where people have designed a system which intelligently identifies malicious domains, analyze them and provide defense information[4]. Some tools like EXPOSURE[5] are also present which leverage machine learning algorithms to tackle malicious domains. In almost every research and article, people agree that days when internet was safe is long gone. This project identifies which countries are most attacked by malicious activities where above stated studies can be effectively implemented to tackle malicious activities.

4 Data Model

Data model of the project is in start schema having 4 dimensions and one fact table. The 4 dimension tables are: DIM_IP, DIM_LOCATION, DIM_CATEGORY, DIM_DATE. All the dimension tables are having primary keys which are being referenced by the foreign keys present in the fact table. DIM_IP holds all the malicious IPs along with few details like its ISP, Domain and Usage Type. Unique Identifier, IP_ID(also the primary key) is also present in this table. IP, ISP, Domain and Usage Type have datatype VARCHAR(50), VARCHAR(100), VARCHAR(100) and VARCHAR(100) respectively with IP_ID of integer datatype. DIM_IP is connected to fact table by IP_ID. Data present in the data model is mainly on IP and Country level and this project is build on analysis of malicious IPs. So, DIM_IP is important to be in the model. DIM_IP is built from the first data source i.e. data coming from `https://www.abuseipdb.com`

DIM_LOCATION is having 11 attributes in it, namely, LOC_ID, REGION, COUNTRY, CITY, Lowest Rate of Malware Infection, Highest Rate of Malware Infection, Highest Number of Users Attacked by Ransomware, Highest Percentage of DDoS Attacks, Least Amount of Freedom on Net, Highest Amount of Freedom on Net, Most Vulnerable Countries to Hacking with data types INTEGER, VARCHAR(50), VARCHAR(50), VARCHAR(50), VARCHAR(10), VARCHAR(10), VARCHAR(10), VARCHAR(10), VARCHAR(10), VARCHAR(10), VARCHAR(10) respectively. DIM_LOCATION is connected to the fact table by LOC_ID. DIM_LOCATION is built from the first data source, second data source i.e. data coming from `https://www.internetworldstats.com` and data coming from `https://cdn.comparitech.com/wp-content/uploads/2017/02/Countries-that-are-`
`jpg`. Region column from the internetworldstats is combined with country and city in DIM_LOCATION. DIM_LOCATION is used for answering the BI queries.

DIM_DATE is having 6 attributes - DATEKEY, DATE, DAYOFMONTH, MONTH, QUARTER AND YEAR with datatype INTEGER, DATETIME, INTEGER, INTEGER, INTEGER, INTEGER respectively. DIM_DATE is built by combining firstreported date and lastreported date present in MALICIOUSIP staging table and then inserting them in DIM_DATE. DIM_DATE is not a very important dimension of the data model currently, but can later be used to show trends of IP reports. DIM_DATE is connected to fact table by DATEKEY.

DIM_CATEGORY have 3 columns in it - CAT_ID, CATEGORY and CATEGORY DESCRIPTION with datatype INTEGER, VARCHAR(max)and VARCHAR(max) respectively. DIM_CATEGORY is included in the model for category attribute. If there is a

confusion regarding any category then categorys' description can be seen in DIM_CATEGORY table. DIM_CATEGORY is connected to fact table by CAT_ID. DIM_CATEGORY is built by using only the first data source `https://www.abuseipdb.com`.

IP_FACT_TABLE is having 9 attributes - IP_ID, LOCATION_ID, CATEGORY_ID, FIRST_REPORTED_DATEKEY, LAST_REPORTED_DATEKEY, TIMES_CATEGORY_REPORTED TIMES_REPORTED, INTERNET_USERS, PERCENT_INTERNET_USERS with datatype INTEGER for all attributes. the fact is references to all dimension tables by foreign keys. It contains all 3 required measures in it which are further used to build cube.

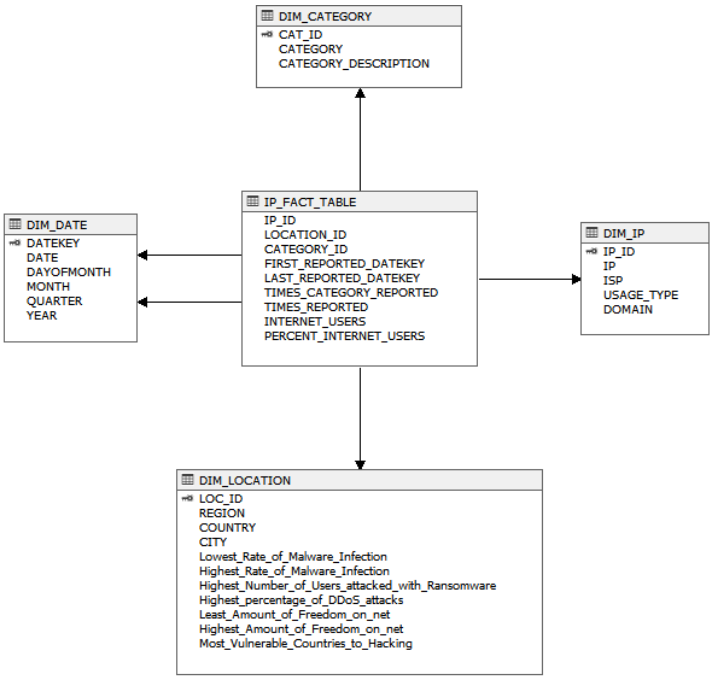

Figure 1: Some star schema i found online

5 Logical Data Map

In this section, describe your logical data map, i.e. how every row of every data source is handled such that it is a part of your star schema.

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 1

Source	Column	Destination	Column	Type	Transformation
abuseipdb.com	IP Address	DIM_IP	IP	Dimension	Extracting IP from url in R - substr(ip_link,33,str_length(ip_link))and
abuseipdb.com	Internet Service Provider	DIM_IP	ISP	Dimension	Table transposed and data arranged in R and sorted in SSIS
abuseipdb.com	Usage Type	DIM_IP	USAGE_TYPE	Dimension	Table transposed and data arranged in R and sorted in SSIS
abuseipdb.com	Domain Name	DIM_IP	DOMAIN	Dimension	Table transposed and data arranged in R and sorted in SSIS
abuseipdb.com	Country	DIM_LOCATION	COUNTRY	Dimension	Table transposed and data arranged in R, only countries having map
abuseipdb.com	City	DIM_LOCATION	CITY	Dimension	Table transposed and data arranged in R
Compritech.com	Lowest Rate of Malware Infection	DIM_LOCATION	Lowest_Rate_of_Malware_Infection	Dimension	Data arranged in excel in required format and 'Y', 'N' flags assigned
Compritech.com	Highest Rate of Malware Infection	DIM_LOCATION	Highest_Rate_of_Malware_Infection	Dimension	Data arranged in excel in required format and 'Y', 'N' flags assigned
Compritech.com	Highest Number of Users attacked with Ransomware	DIM_LOCATION	Highest_Number_Of_Users_attacked_with_Ransomware	Dimension	Data arranged in excel in required format and 'Y', 'N' flags assigned

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
Compritech.com	Highest percentage of DDoS attacks	DIM_LOCATION	Highest_percentage_of_DDoS_attacks	Fact	Data arranged in excel in required format and 'Y', 'N' flags assigned
Compritech.com	Least Amount of Freedom on net	DIM_LOCATION	Least_Amount_of_Freedom_on_net	Dimension	Data arranged in excel in required format and 'Y', 'N' flags assigned
Compritech.com	Highest Amount of Freedom on net	DIM_LOCATION	Highest_Amount_of_Freedom_on_net	Dimension	Data arranged in excel in required format and 'Y', 'N' flags assigned
Compritech.com	Most Vulnerable Countries to Hacking	DIM_LOCATION	Most_Vulnerable_Countries_to_Hacking	Dimension	Data arranged in excel in required format and 'Y', 'N' flags assigned
abuseipdb.com	Category	DIM_CATEGORY	CATEGORY	Dimension	No Transformation
abuseipdb.com	Category Description	DIM_CATEGORY	CATEGORY_DESCRIPTION	Dimension	No Transformation
abuseipdb.com	Date	DIM_DATE	DATE	Dimension	No Transformation
abuseipdb.com	Day of Month	DIM_DATE	DAYOFMONTH	Dimension	DATEPART(DAY, DATE)
abuseipdb.com	Month	DIM_DATE	MONTH	Dimension	DATEPART(MONTH, DATE)
abuseipdb.com	Quarter	DIM_DATE	QUARTER	Dimension	DATEPART(QUARTER, DATE)
abuseipdb.com	Year	DIM_DATE	YEAR	Dimension	DATEPART(YEAR, DATE)
abuseipdb.com	Number of times category reported recently	IP_FACT_TABLE	TIMES_CATEGORY_REPORTED	Fact	Group_by(Categories) and summarize(n()) in R

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
abuseipdb.com	Number of times IP has been reported	IP_FACT_TABLE	TIMES_REPORTED	Fact	(CASE WHEN ROW_NUMBER() OVER (PARTITION BY A.IP O
internetworldstats.com	Internet Users	IP_FACT_TABLE	$INTERNET_USERS$	Fact	(CASE WHEN ROW_NUMBER() OVER (PARTITION BY A.C C.INTERNET_USERS ELSE 0)
internetworldstats.com	Percentage of Internet Users	IP_FACT_TABLE	$PERCENT_INTERNET_USERS$	Fact	(CASE WHEN ROW_NUMBER() OVER (PARTITION BY A.C C.PERCENT_INTERNET_USERS ELSE 0)

6 ETL Process

The ETL process in SSIS is divided in 6 stages. First stage executes the R code which is used to get data from `https://www.abuseipdb.com`. Two R Codes creating 3 CSV files are executed in this stage. This stage was the most challenging stage because data was fetched from a table present on website which didn't have any headers and the actual attributes of the table were present in a single column in multiple rows. Took time to figure out the no header problem. Another challenge was that there was a HOSTNAME field present for most of the IPs in the same fashion like other attributes were present in multiple rows. But, for few IPs, this field was not present which another major challenge to extract data for all fields except the HOSTNAME.

Second stage checks the existence of all the staging tables. All staging tables in this stage are truncated except the Malicious IP staging table and staging table for category wise count. These two tables are not truncated because data is being appended in these 2 tables. Appending data in these two tables was needed because the R code is taking too long to execute for all IPs (around 1200) for a single page. So, R Code is executed in batches of 20-30 at once and then data is appended in the staging table. In the third stage, all the staging tables are populated by pulling all relevant flat files and then dumping those files in staging tables. In this stage, errors like - 'Data Truncated' were faced because the size of destination column of staging table was less as compared to the size of input data.

Fourth stage is where the fact table is dropped, so as to truncate the dimension tables which would not be possible if fact is not dropped because of primary-foreign key relationship. In fifth stage, all the dimension tables are truncated and then an insert query is executed to populate all dimension tables. Here too, problem like 'Data Truncated' was faced. Finally, in the sixth stage, fact table is populated.

7 Application

(Req-1) Relationship of country wise internet users with number of IP reports – It has been shown by making a country wise bar graph with side by side bars comparing and showing relationship between 'Number of times an IP is Reported' and 'Percentage of Internet Users' for each country

(Req-2) Relationship of countries having "Highest Number of Ransomware Attacks" with IP reports – Filter has been applied to 'Y' for "Highest Number of Ransomware Attacks" and the pie chart then shows country wise figure of how many times IPs have been reported.

(Req-3) Relationship of Categories under which IPs have been recently reported to countries with "Highest Rate of Malware Infection" – Filter for "Highest Rate of Malware Infection" has been put to 'Y' which filters all records falling under "Highest Rate of Malware Infection". Then a line chart has been built for seeing the trend w.r.t multiple categories by number of times an IP has been reported under that category.

7.1 BI Query 1: Which country is having least number of IP reports as compared to proportion of population using Internet in Europe

For this query, the contributing sources of data are: Source 1: Abuseipdb.com and Source 2: internetworldstats.com

We can clearly see that although in Iceland and Sweden around 99 percent and 97 percent of the population is using internet but the number of times any IP has been reported is very less as compared to France where 93 percent of population use the internet, but the number of times an IP is reported is very high. As per current dataset available for analysis it can be said that in Europe, malicious activities occur more in France. as illustrated in Figure 2.

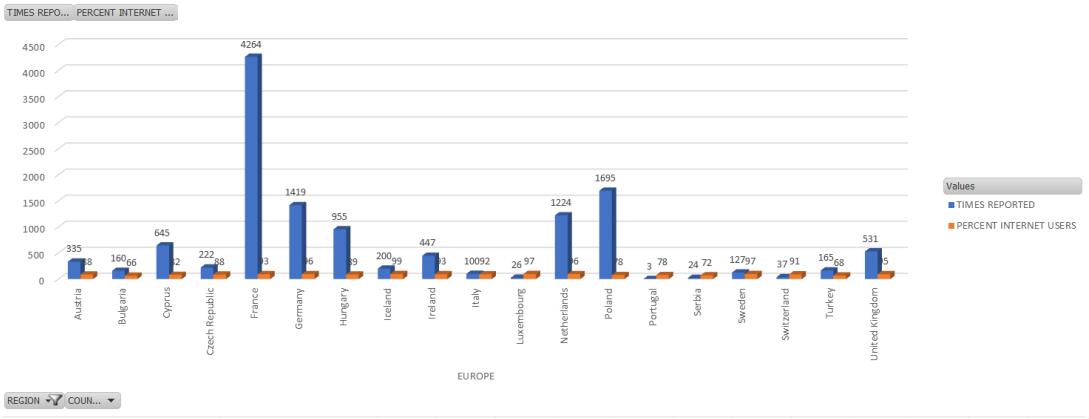

Figure 2: Results for BI Query 1

7.2 BI Query 2: What relationship does the countries having "Highest number of Ransomware Attacks" have with the number of IP reports?

For this query, the contributing sources of data are: Source 1: Abuseipdb.com and Source 3: compritech.com

It can be seen in the pie chart that France falls under highest number of ransomware attacks as illustrated in Figure 3. and also covering the major chunk of pie chart. Based on current dataset loaded for cube, it can be related that risk of cyber threats in France is more as compared to other countries.

7.3 BI Query 3: Under which category IPs have been reported the most in countries with "Highest Rate of Malware Infection"?

For this query, the contributing sources of data are: Source 1: Abuseipdb.com and Source 3: compritech.com

From the line chart we can see that in countries with highest rate of malware infection, most of the times IPs have been reported under the category "Brute-Force SSH" followed by "Port Scan" as illustrated in Figure 4.

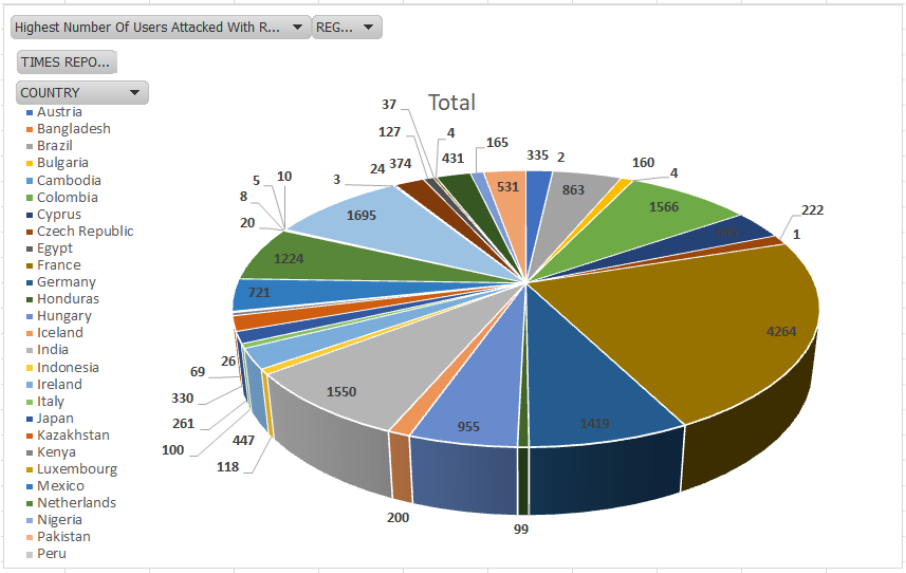

Figure 3: Results for BI Query 2

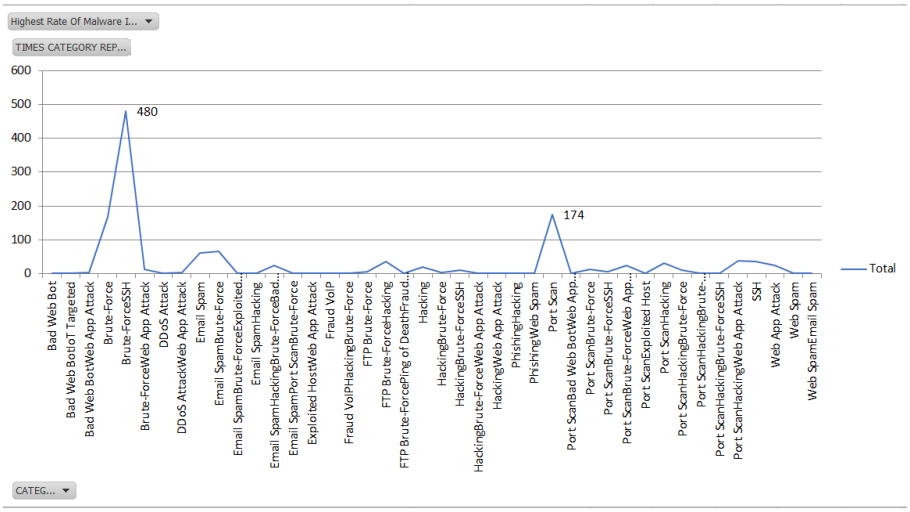

Figure 4: Results for BI Query 3

7.4 Discussion

Summarising the result of above BI queries we can see that in Europe, France is having high number of IP reports and falls under list of countries where highest number of users were attcked by ransonware. Also, as per our current dataset, we found that in countries having highest rate of malware infection "Brute-ForceSSH" is the most report category followed by Port Scan. This clearly means that implementing studies which tackle malicious domains as discussed above in section 3 will control the cyber attacks in France.

8 Conclusion and Future Work

As we know that cyber threats are increasing day by day, countries falling trap to ransomware and other malicious activities are also increasing. Project gives a high level view

of activities related to malicious IPs.

This project can further be extended and and analyze various methods throught which cyber threats are made and check which countries are affected the most.

References

Appendix

[1] Classification of Unknown Web Sites Based on Yearly Changes of Distribution Information of Malicious IP Addresses (2018) 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), New Technologies, Mobility and Security (NTMS), 2018 9th IFIP International Conference on, p. 1. doi: 10.1109/NTMS.2018.8328683.

R code example

```
#-----------------------------------------------Libraries-------------------
library(tidyverse)
library(rvest)
library(stringr)
library(htmltab)
#-----------------------------------------------Main URL--------------------
url <- 'https://www.abuseipdb.com/sitemap'
#--------------------Functions to fetch Malicious IPs and their urls(to get
IP_function <- function(html){

  IP <- html %>% html_nodes('.col-md-2␣a') %>% html_attr("href") %>% str_tri

  trimws(IP) %>% str_replace_all("\\n", "")

  a <- IP[str_length(IP) <= 47]

}
url_data <- function(url){

  html <- read_html(url)

  IP_function(html)
}
#-------------------------------------Functions to fetch Data for each IP----
table_url_data <- function(ip_link){

  html <- read_html(ip_link)

  table <- table_function(html)

  IP <- substr(ip_link,33,str_length(ip_link))

  a <- as.tibble(cbind(IP, table))
  #tibble(IP=IP, table)
}
```

```
table_function <- function(html){

  table <- html %>%

    html_nodes(xpath='//*[@id="report-wrapper"]/div[1]/div[1]/div/table') %>

    html_table() %>% as.data.frame()

  table <- t(table)

  table <- as.data.frame(table)

  colnames(table) <- as.character(unlist(table[1,]))

  table <- table[-1,]

  if (ncol(table) == 6) {
    table <- table[,-3]
  }

  table <- as.tibble(table)
}
#----------------------------------Functions to fetch Number of Times Reporte
reported_url_data <- function(url){

  html <- read_html(url)

  reported <- reported_function(html)

  IP <- substr(url,33,str_length(url))

  tibble(IP=IP, Reported=reported)
}
reported_function <- function(html){

  reported <- html %>%

    html_nodes('#report+␣p␣b:nth-child(1)') %>%

    html_text() %>%

    str_trim() %>%

    unlist()

  if_else(is.na(reported), '0', reported)
}
#----------------------------------Functions to fetch First Reported Date--
firstreported_url_data <- function(url){

  html <- read_html(url)
```

```
  firstreported <- firstreported_function(html)

  IP <- substr(url,33,str_length(url))

  firstreported <- substr(firstreported,1, 10)

  tibble(IP=IP, FirstReported=firstreported)
}
firstreported_function <- function(html){

  firstreported <- html %>%

    html_nodes('b+ time') %>%

    html_attr("datetime") %>%

    str_trim() %>%

    unlist()

  #if_else(is.na(firstreported), "Not Found", firstreported)

}
#-------------------------------------Functions for Last Reported Date-----
lastreported_url_data <- function(url){

  html <- read_html(url)

  lastreported <- lastreported_function(html)

  IP <- substr(url,33,str_length(url))

  lastreported <- substr(lastreported,1, 10)

  tibble(IP=IP, Lastreported=lastreported)
}
lastreported_function <- function(html){

  lastreported <- html %>%

    html_nodes('b time') %>%

    html_attr("datetime") %>%

    str_trim() %>%

    unlist()

  #if_else(is.na(lastreported), "Not Found", lastreported)
}
```

```r
#-------------------------------------------------------------Comment Category
commentcategory_url_data <- function(url){

  commentcategory <- htmltab(doc=url, which=2)

  commentcategory <- commentcategory %>% group_by(Categories) %>% summarize(

  IP <- substr(url,33,str_length(url))

  commentcategory_IP <- cbind(IP, commentcategory)

  as.tibble(commentcategory_IP)
}
#--------------------------------Main function which calls all other functio
data_table <- function(url, i){

  all_ips <- lapply(str_c(url,"?page=",1:i),

                    function(url){

                      all_ips <- url %>% url_data()
                    })

  ip_link <- unlist(all_ips)

  ip_link <- ip_link[1:10]

  for (i in ip_link) {

    table_data <- ip_link %>% map(table_url_data) %>% bind_rows()

    REPORTED <- ip_link %>% map(reported_url_data) %>% bind_rows()

    FIRSTREPORTED <- ip_link %>% map(firstreported_url_data) %>% bind_rows()

    LASTREPORTED <- ip_link %>% map(lastreported_url_data) %>% bind_rows()

    commentcategory <- ip_link %>% map(commentcategory_url_data) %>% bind_ro

  }

  ISP_USAGE_DOMAIN_COUNTRY_CITY_REPORTED <- merge(table_data, REPORTED)

  ISP_USAGE_DOMAIN_COUNTRY_CITY_REPORTED_FIRSTREPORTED <- merge(ISP_USAGE_DO

  ISP_USAGE_DOMAIN_COUNTRY_CITY_REPORTED_FIRSTREPORTED_LASTREPORTED <- merge

  write.csv(ISP_USAGE_DOMAIN_COUNTRY_CITY_REPORTED_FIRSTREPORTED_LASTREPORTE

  write.csv(commentcategory, "Category␣Wise␣Count.csv", row.names = F)
```

```
}
#----------------------------------------Calling Main Function-------------
data_table(url,1)
```

[2] Jakalan, A. et al. (2016) Social relationship discovery of IP addresses in the managed IP networks by observing traffic at network boundary, Computer Networks, 100, pp. 1227. doi: 10.1016/j.comnet.2016.02.012

[3] Hubballi, N. and Tripathi, N. (2017) An event based technique for detecting spoofed IP packets, Journal of Information Security and Applications, 35, pp. 3243. doi: 10.1016/j.jisa.2017.04.001.

[4] Hubballi, N. and Tripathi, N. (2017) An event based technique for detecting spoofed IP packets, Journal of Information Security and Applications, 35, pp. 3243. doi: 10.1016/j.jisa.2017.04.001.

[5] https://www.cs.ucsb.edu/ chris/research/doc/tissec14$_e$*xposure.pdf*