Faculty of Technology, Natural Sciences and Maritime Sciences

Data Analysis and Visualization

Rahmat Mozafari

November 2, 2020

Contents

1 Introduction **5**

2 Methods **5**

3 Tools **6**

4 Results **8**
4.1 Geolocation of IP addresses 8
4.2 Visualization 12
4.3 Bad IPs 26
4.4 Database 30

5 Conclusion **32**

References **33**

List of Tables

1 Shows the number of counties and the number of IP's 8
2 Shows top five ip's with high activity 8
3 Shows top five countries 8
4 Threat types description 17

List of Figures

1 Data analysis and visualization 7
2 Shows the geolocation of IP addresses 10
3 Shows the geolocation of users based on continents and cities . . 11
4 Shows activity on the server each day 12
5 Shows activity for each IP 13
6 List of continents 14
7 Country in, Africa, Europe, Asia and South America 15
8 Country in North America and in Oceania 16
9 Shows perform 16
10 Shows threat types 17
11 Shows actions and threat types by date 18
12 Shows threat's continents and countries 19
13 Shows threats from Europe continent by dates 20
14 Shows threats from Asia, South and North America 21
15 Shows attacker with country names 22
16 Shows attacker from Europe, North America and Asia, groupped by perform on the server-side 23
17 Shows abuser with country names 24

18 Shows abuser from Europe, Asia and North America 25
19 Shows IPs which is threat 27
20 Shows IPs which is known abuser 27
21 Shows the same IPs which is threat, is tor network user and is knwon attacker 28
22 Lits of bad IPs 29
23 Database Schema 31

Abstract

In this paper, we will analyze and visualize the data we have, and based on this information, we will find IP addresses that will pose a danger in the future. We will analyze our data based on the continent, country, and city and we will analyze them clearly with the help of graphs and diagrams. From the information we obtained through analysis, we concluded that IP addresses from some countries cause security problems on the server. We think that it will be very important for the security of the server to be careful with the IP addresses in the table we have given at the end of the project.

1 Introduction

To make this project, we had more than 71000 IP addresses and server data related to them, which were formed in 5 log documents. We aimed to first access the detailed location data of these IP addresses and analyze and visualize these data by processing them. However, with these analyzes and visualizations, it is to find the factors that will create a danger factor from this information and maybe to take this to an advanced level and to make a quick decision about each visitor entering the server by creating a Machine Learning model. As we know, data security is very important to us, we need to protect it from data thieves. Therefore, we must thoroughly analyze our server visit data, and with that, we need to make a quick decision about visitors.

2 Methods

When we got our hands on this project, we had 5 log documents and over 71000 data sets consisting of them. The first thing we would do was extract this data. Because the first and most important step of the data analysis process was to retrieve the data, find incomplete or error-processed data and analyze them, if this data does not affect the analysis we will do, to extract it. The first processing of our data took us a little while, and the next step we dealt with was how to access location information based on the IP addresses we had. To conclude this step, we tried a few solutions.

1. Python's Selenium package: This solution worked, but it was slow. Because he has to open the Web Browser once each time to get the location information. So this was a very slow solution.

2. Getting location information with Bash script in Linux: In this solution, we wrote a Bash script and obtained the location information with the help of the location API. This solution was much faster than the previous solution, but not as much as we wanted.

3.Using the Python package prepared by companies that share location information: In this solution, we obtained all the necessary location information at the speed we wanted by using the Python package of the company that provides location information, but since this was a fee, it had a daily limitation of 1500. For this reason, we finished this process in 2-3 days.

After obtaining the location information, we preprocessed this data and made it usable by extracting the outliers and missing data. While doing these operations, we used Python's Pandas, NumPy, and Seaborn libraries, and here we made all our data ready for use in Database.

While creating the database, we divided our data into a few categories. These

are Continent, Country, Server_Info, Threat ... etc and while doing these, we gave importance to processes such as data redundancy and normalization.

We tried to learn all the necessary libraries of Python to be able to analyze and visualize data for this project. We understood how important Pandasin and Seaborn libraries are in preprocessing and visualizing data, and understood the limitation that the same SQL operations can be done with Python for the first time. But since Python and SQLs have their good and bad sides, we can do a lot of things in the data world by using them together.

We tried to learn all the necessary libraries of Python to be able to analyze and visualize data for this project. We understood how important Pandasin and Seaborn libraries are in preprocessing and visualizing data, and understood the limitation that the same SQL operations can be done with Python for the first time. But since Python and SQLs have their good and bad sides, we can do a lot of things in the data world by using them together.

Maybe if we had enough time, we'd have to dig deeper into our Data. For example, we would analyze data on a city basis. Because we have 1132 unique city information, we can't analyze and visualize them in a short time.

3 Tools

To visualize the data, we used Python libraries such as Pandas, Seaborn, MatplotPy, and the Microsoft Power BI tool. My personal view is that Python's libraries are very powerful in data processing and data visualization, and you can get the statistics you want according to the data, and you can group and visualize according to the variables you want. And in data visualization, another tool where you can get the same result without much coding is Power BI and Tableau. For these reasons, we used Power BI in this project while visualizing some data. We also used postgreSQL and DBeaver for our database. For the geoinformations we used API from ipdata.com and we also used a free version database from maxmind.com.

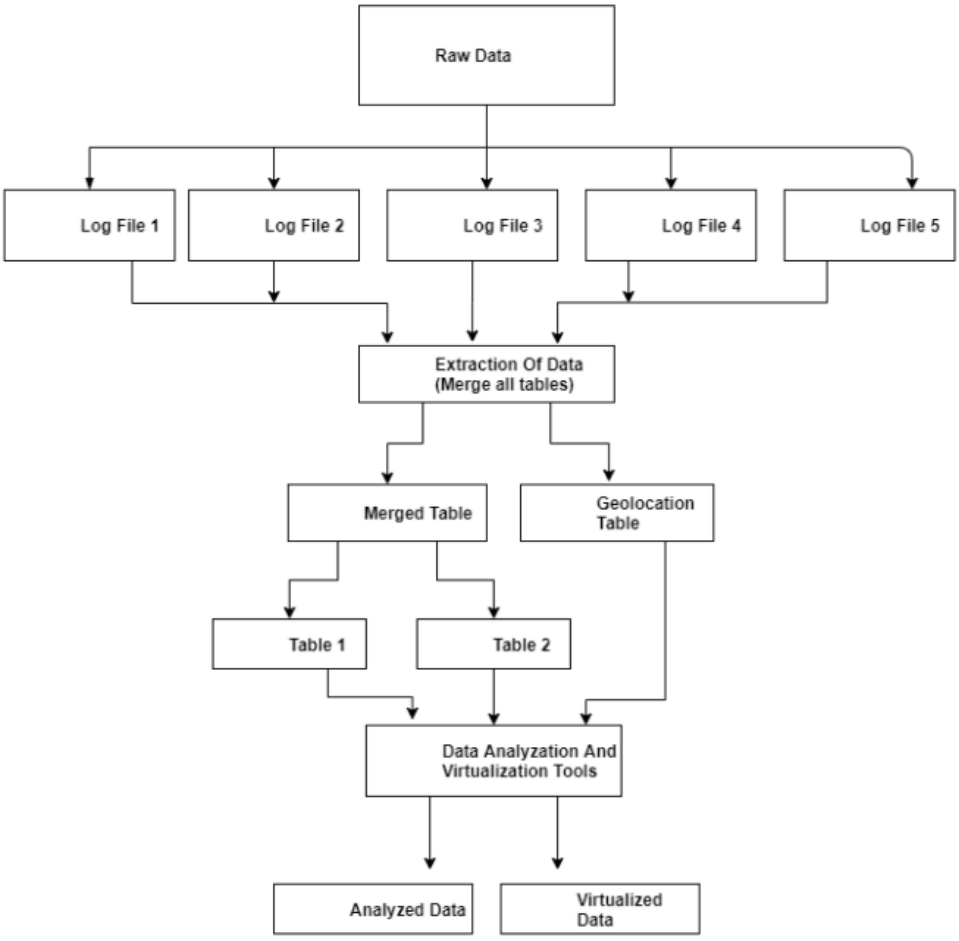


Figure 1: Data analysis and visualization

4 Results

Finally, in this section, we are walking through our results, which we gathered through our study. After we sorted and organized the data, we started to analyze it. There are 128 countries and 6563 different IP addresses. We have located each IP address's geolocation and tried to find as much information about each IP as possible. Some users have very high login activity, and also some countries have more than 10000 requests on the server. We created some tables to show the total number of countries, the total number of users, and the top five users, and the top five countries with high request on the server.

Number of counties	Number of IP's
128	6563

Table 1: Shows the number of counties and the number of IP's

IP addresses	Number of activity
112.85.42.195	429
112.85.42.172	404
49.88.112.73	379
222.186.175.216	370
112.85.42.173	369

Table 2: Shows top five ip's with high activity

Country	Number of request
China	31984
United States	11178
Republic of Lithuania	3698
France	3001
Singapore	1853

Table 3: Shows top five countries

4.1 Geolocation of IP addresses

To find each IP address's geolocation, we have first used [Max] to get geolocation, but the accuracy of latitude and longitude on the country level is 98%, on the state level is 80% and on the city level is 68%. However, after our first feedback, we decided to use [ipd] since we needed to collect some more information about each IP addresses; for instance, if an IP is anonymous or is known as an attacker, then it is a bad IP address and should be blocked to access some information in from the server. The ipdata-APIs accuracy of latitude and longitude is often the near center of the population. These APIs has its advantages and

disadvantage. The downside of ipdata APIs is that it limits how many requests we can send per day. The benefit is that it is free and easy to use.Figure [2] show us the geolocation of each IP addresses based on their latitude and longitude. We separated the user's geolocation into continents and also into cities, which we can see in figure [3]. There are in total 1132 cities, and we have 6563 unique users, but in general, they had 71,824 login activities during these 33 days.

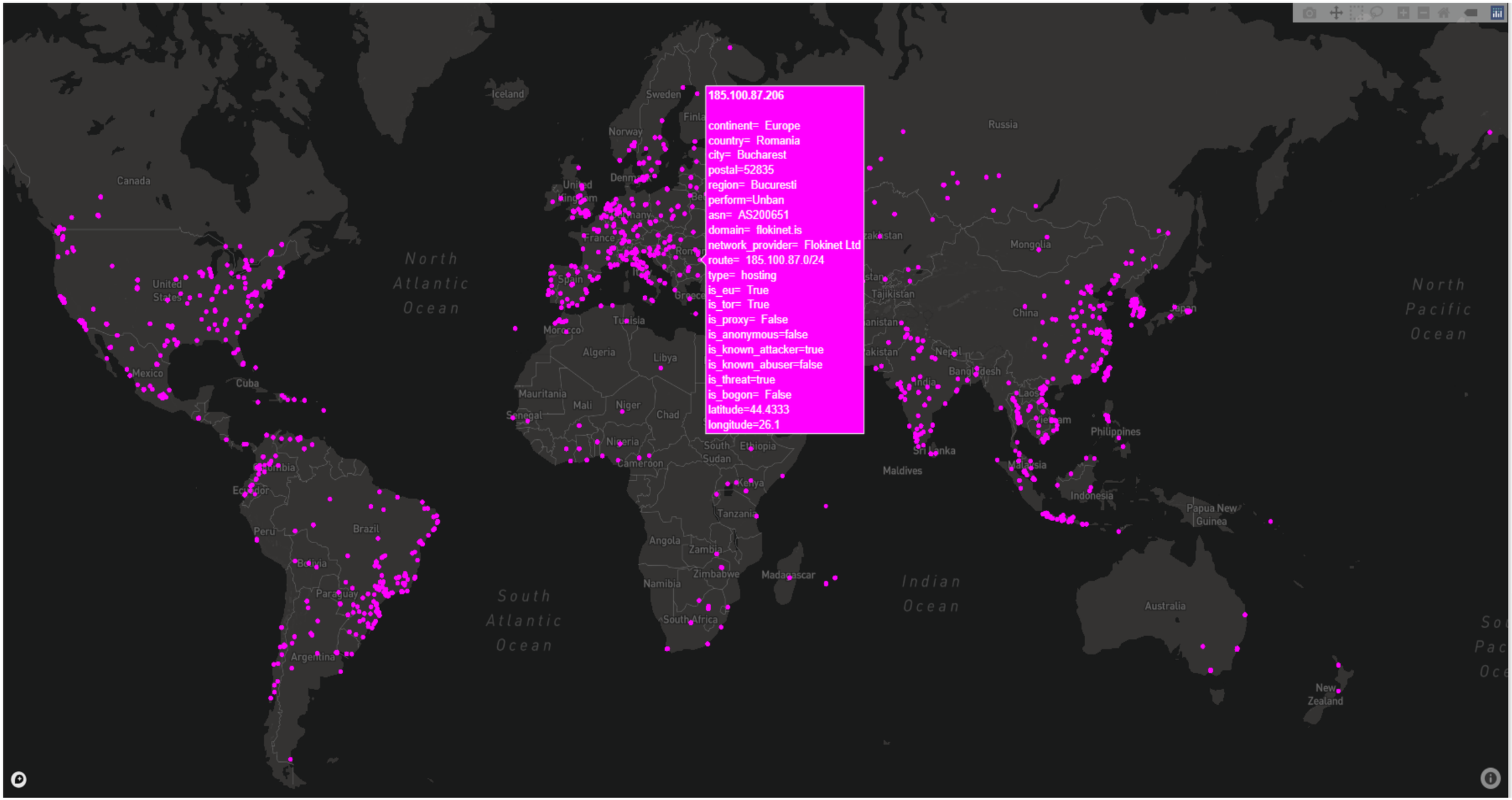


Figure 2: Shows the geolocation of IP addresses

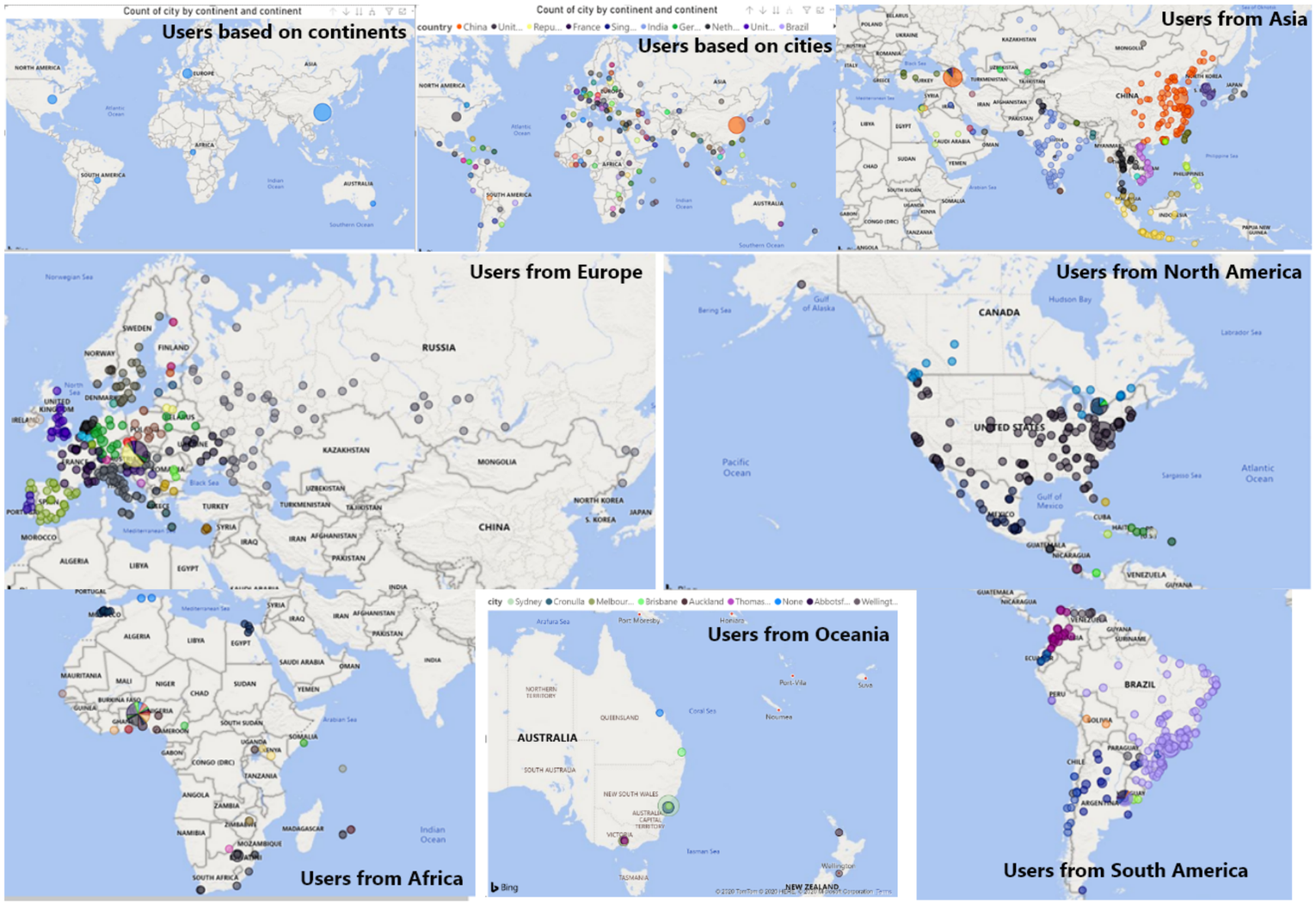


Figure 3: Shows the geolocation of users based on continents and cities

4.2 Visualization

In this our subsection, we are going to presenting our the data we have visualized and analyzed. Firstly, we want to show the daily based activity on the server in figure[4]. The x-axis shows dates, and the y-axis shows the number of activities per day on the server. The activity varies daily, and some days the server goes warm of tasks. However, if the server is small and might go down sometimes when many requests come in, it will not handle it.

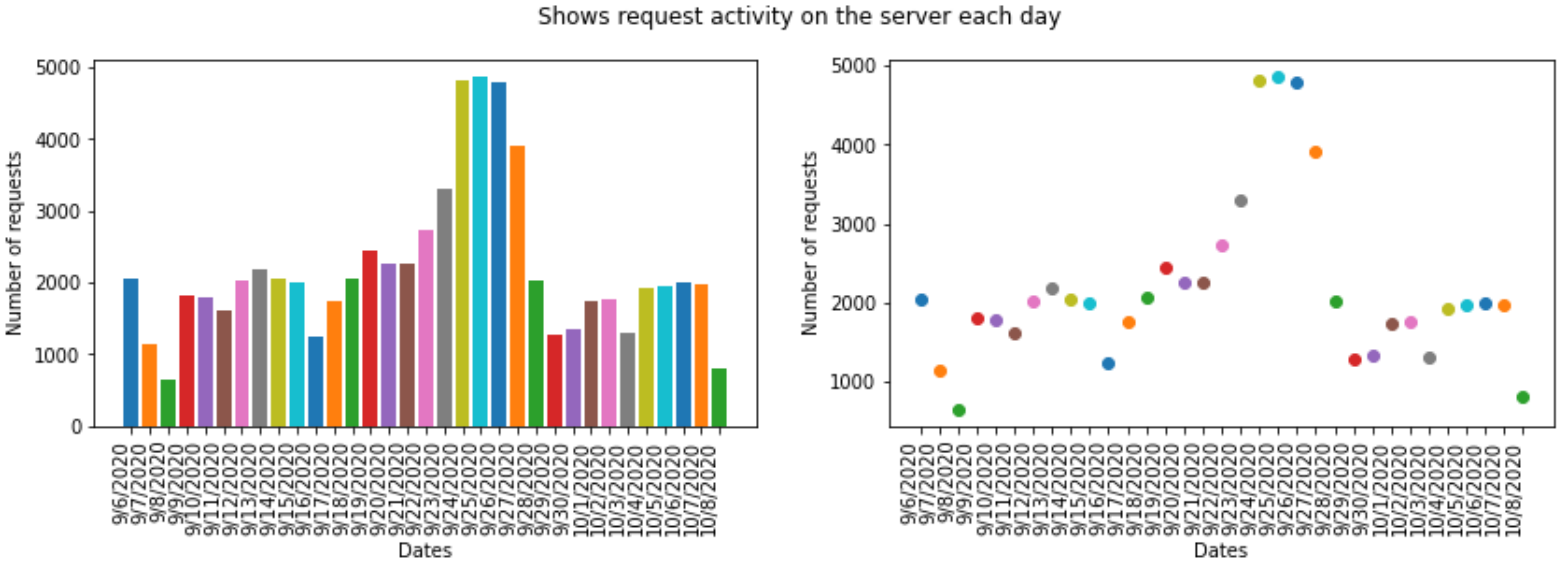


Figure 4: Shows activity on the server each day

The figure[5] describes the request activity for each IP address on the server during the 33 days. In this case, we have grouped the user's activity into six groups. Group 1-4 shows the users who sent 1-40 logins requests to the server during these 33 days. However, on the other hand, group 5 shows the users who sent 1-100 logins requests, and group 6 the users distinct from all other groups because they have sent 1-429 logins requests during these 33 days.

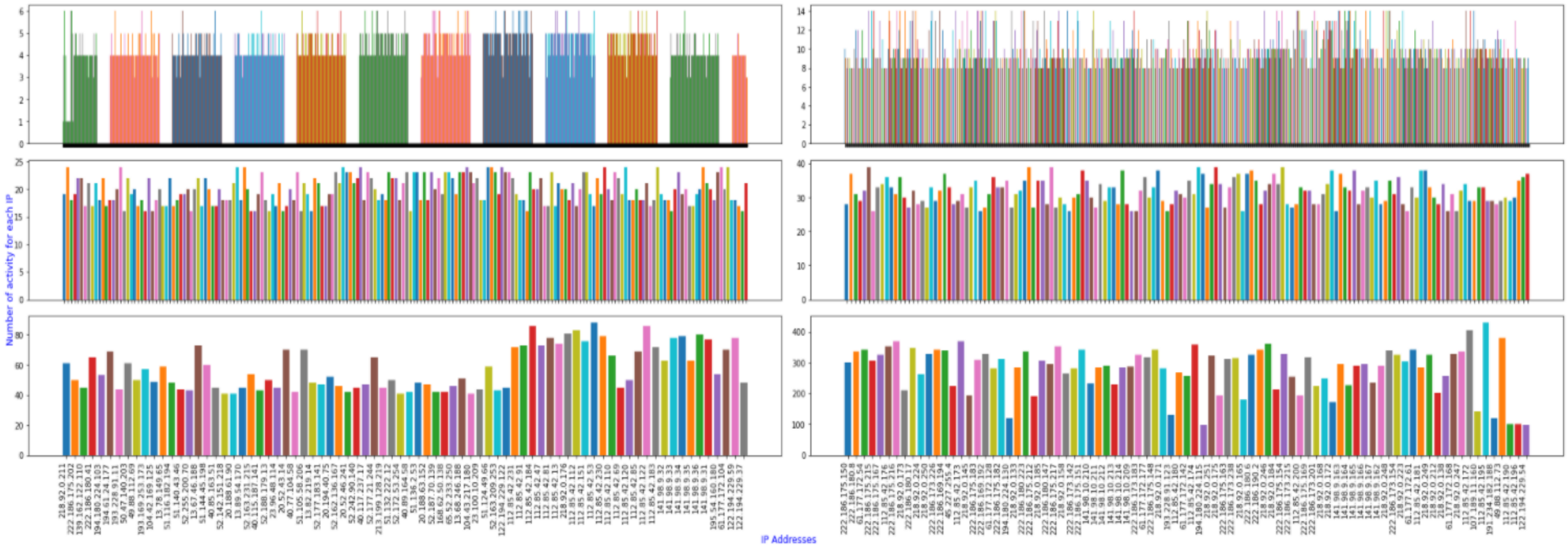


Figure 5: Shows activity for each IP

We have categorized the logins requests into continents based on the user's geolocation. Figure [6] shows how many logins requests come from which country. There are six continents, and Asia is the continent that stands out from other continents when it comes to logins requests. We have totally found 71824 IPs information in our server information over 6 different continents and 128 countries. These are separately Africa 637, Asia 41332, South America 2387, North America 12551, Europe 14646 and Oceania 226. Also, there are 34 European countries, 41 Asian countries, 3 Oceanian countries, 10 South American countries, 11 North American countries and 29 African countries.

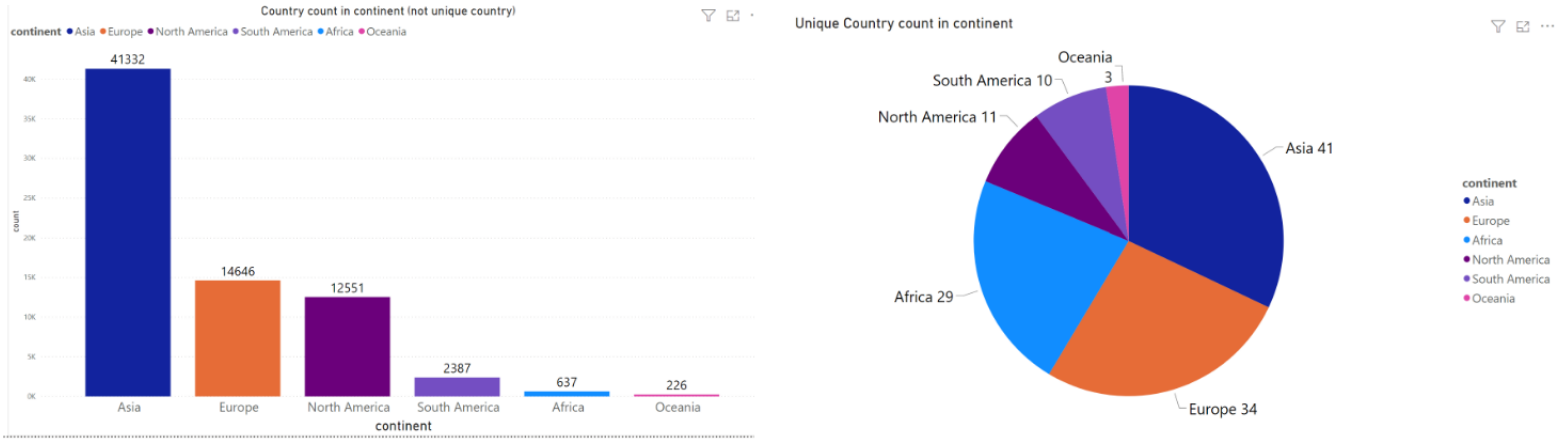


Figure 6: List of continents

At the same time, we obtained 637 IP information from 29 different states from the African continent and the country with the most 165 IP visits is South Africa, 41332 IP information from 41 different states from the Asian continent and the country with the most 31984 IP visits is china, 2387 IP information from 10 different states from the South America and the country with the most 1427 IP visits is Brazil, 12551 IP information from 11 different states from the North America and the country with the most 11178 IP visits is United States, 14646 IP information from 34 different states from the Europe and the country with the most 3698 IP visits is Republic of Lithuania, 226 IP information from 3 different states from the Oceania and the country with the most 208 IP visits is Australia. Figure[7][8] shows this information in bars.

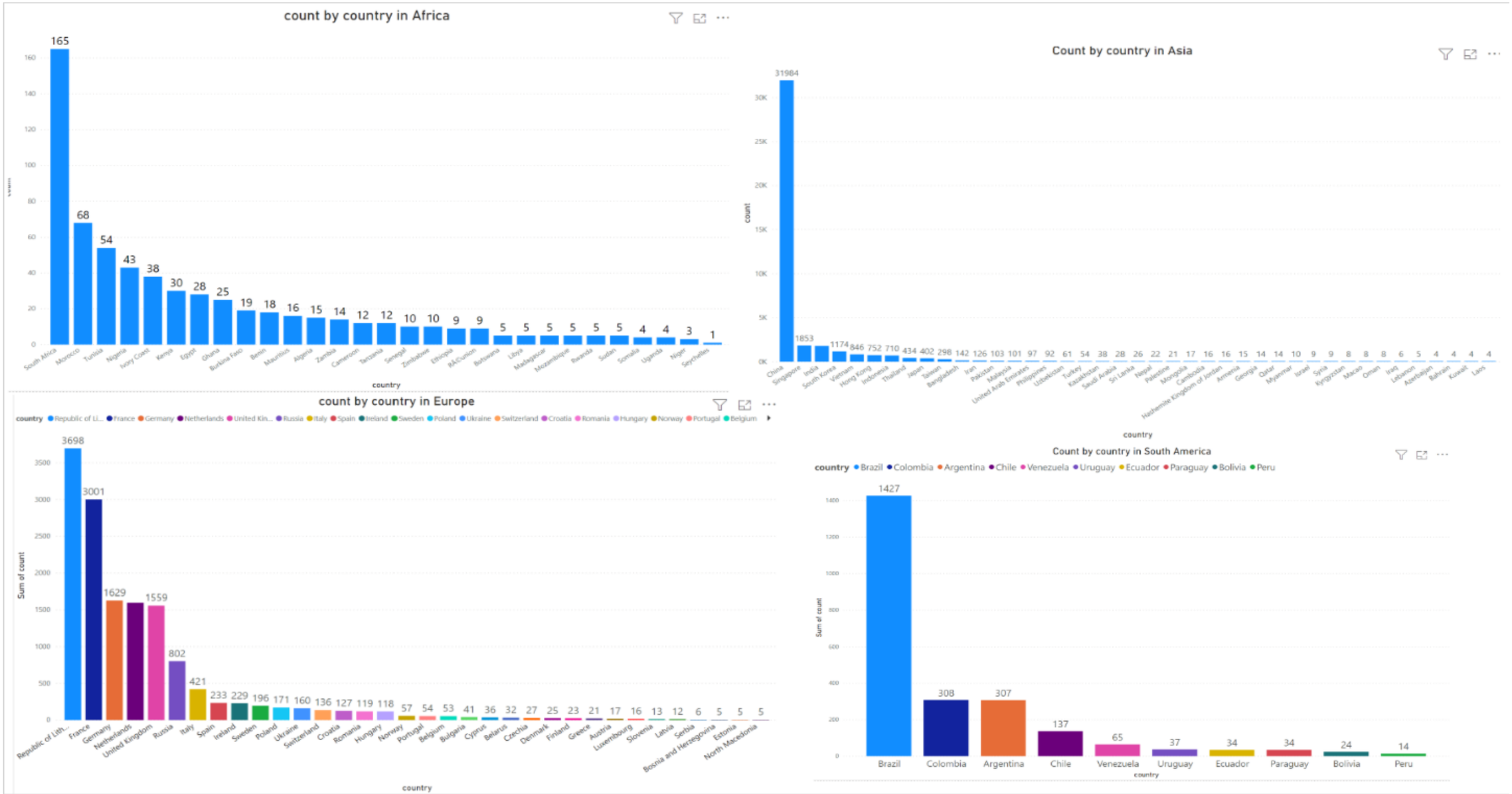


Figure 7: Country in, Africa, Europe, Asia and South America

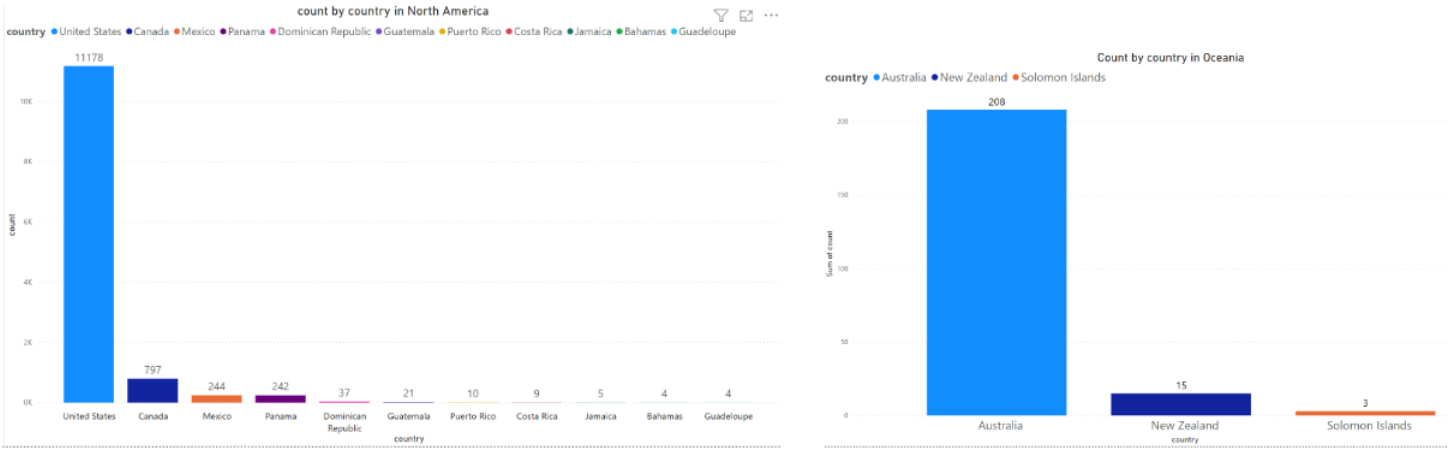


Figure 8: Country in North America and in Oceania

Figure[9] shows the four categories on the server-side and how many of them took place during these 33 days. "Found": are users logged in successfully: "Ban": are users who are banned after they gave an invalid password or username. "Unban": are users who were "Ban" in some hours but logged in successfully after that. "Already banned": are users who were "Ban" and trying to log in, but the "Ban" hours are not finished yet. In other words, it shows how many requests were Found, Ban, Unban, or Already banned.

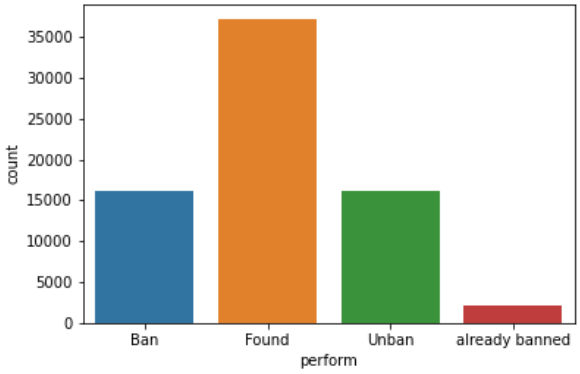


Figure 9: Shows perform

We found a total of 152 unique malicious IP addresses, but they have been in total active 1,282 times on the server during these 33 days. There four types of threats, which are the following:

Types of threat	Explanation
is-tor	is true if the user is associated with a tor network
is-known-attacker	is true if the user is a known source of malicious activity, i.e. malware, attacks etc.
is-knwon-abuser	is true if the user is a known source of abuse, i.e., harvesters, spam etc.
is-threat	is true if one of two is-known-attacker or is-knwon-abuser is true

Table 4: Threat types description

Figure[10] describes the threat types during the 33 days and shows how many of each of them have been. 656 are the same users which used tor network, is a known source of malicious and is threat. There are 142 users which is only threat and 484 users are only abuser. Notice this figure shows a general overview of threats. We haven't removed duplicates of IPs, yet. Figure [11] shows more detail about each threat types.

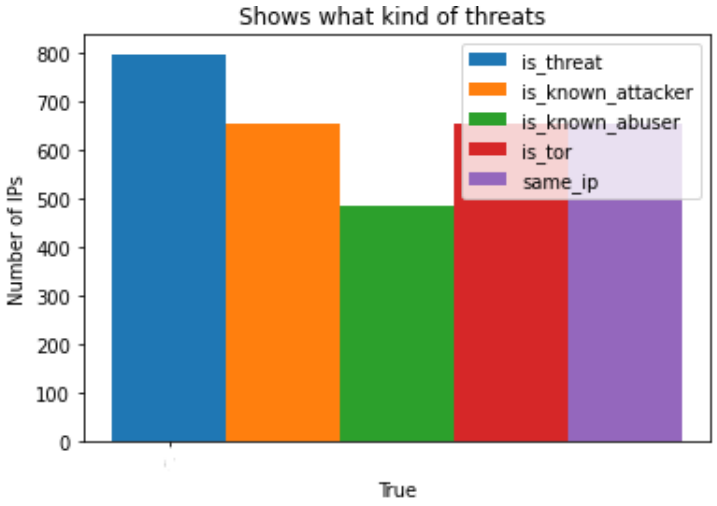


Figure 10: Shows threat types

There are found totally 798 threats from 4 different continents and the continent with highest threat is Europe. 509 threats came from 14 different European countries, and the country with the highest threat is Germany. 51 threats came from 6 different Asian countries, and the country with the highest threat is Singapore. 233 threats came from 2 different North American countries, and the country with the highest threat is USA. 5 threat came from 1 South American country, and the country is Colombia. See figure[12], figure[13] shows threats from only Europe continent by dates and figure[14] shows threats from Asia, North and South Amercia.

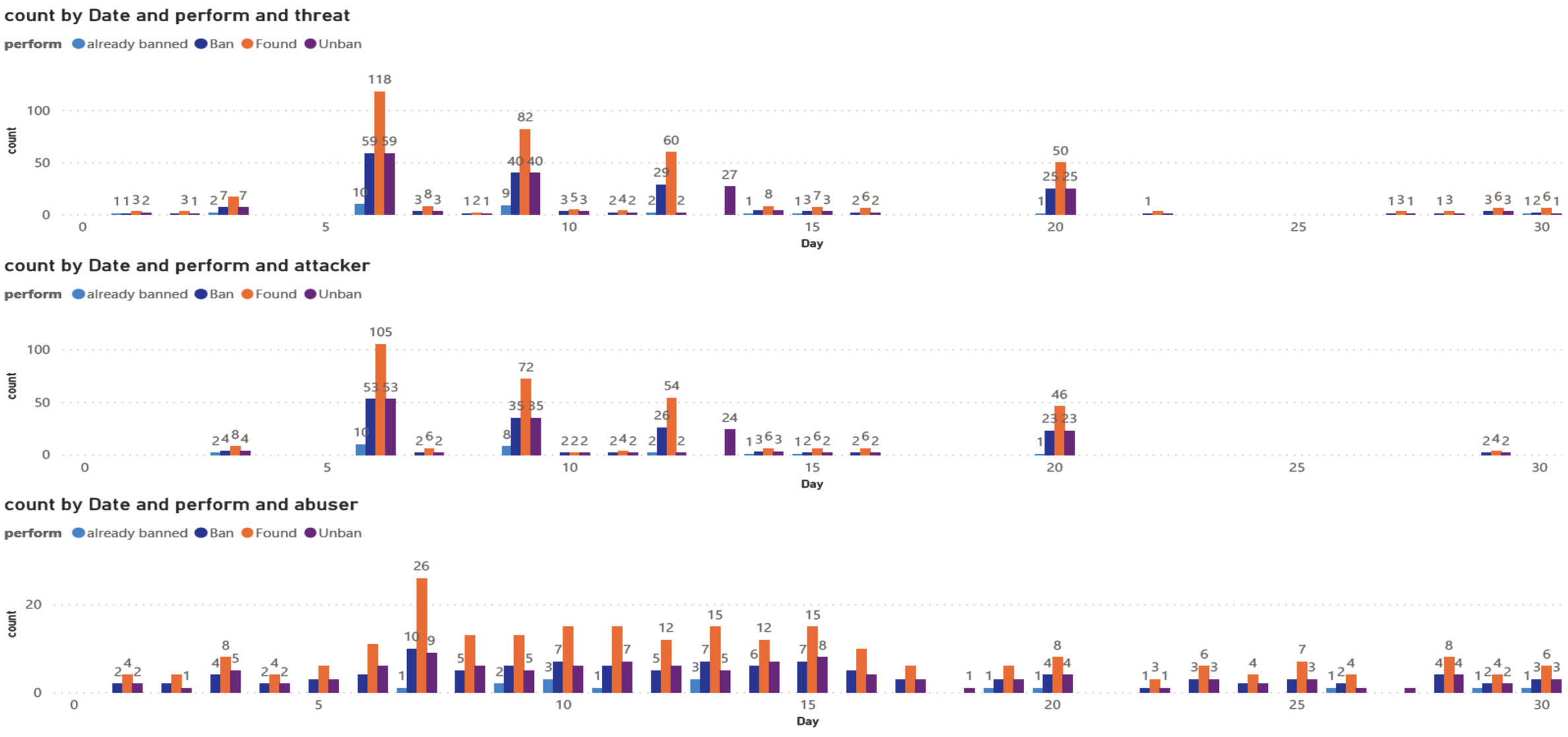
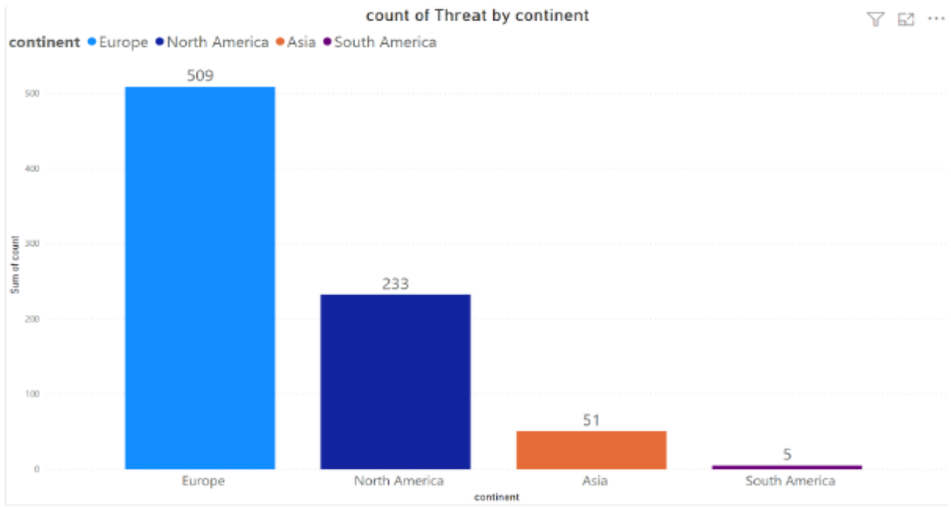
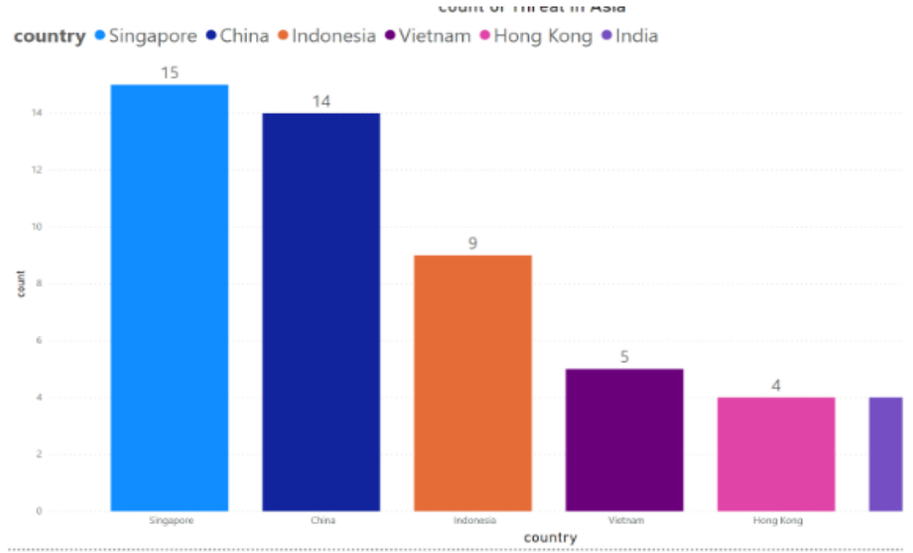
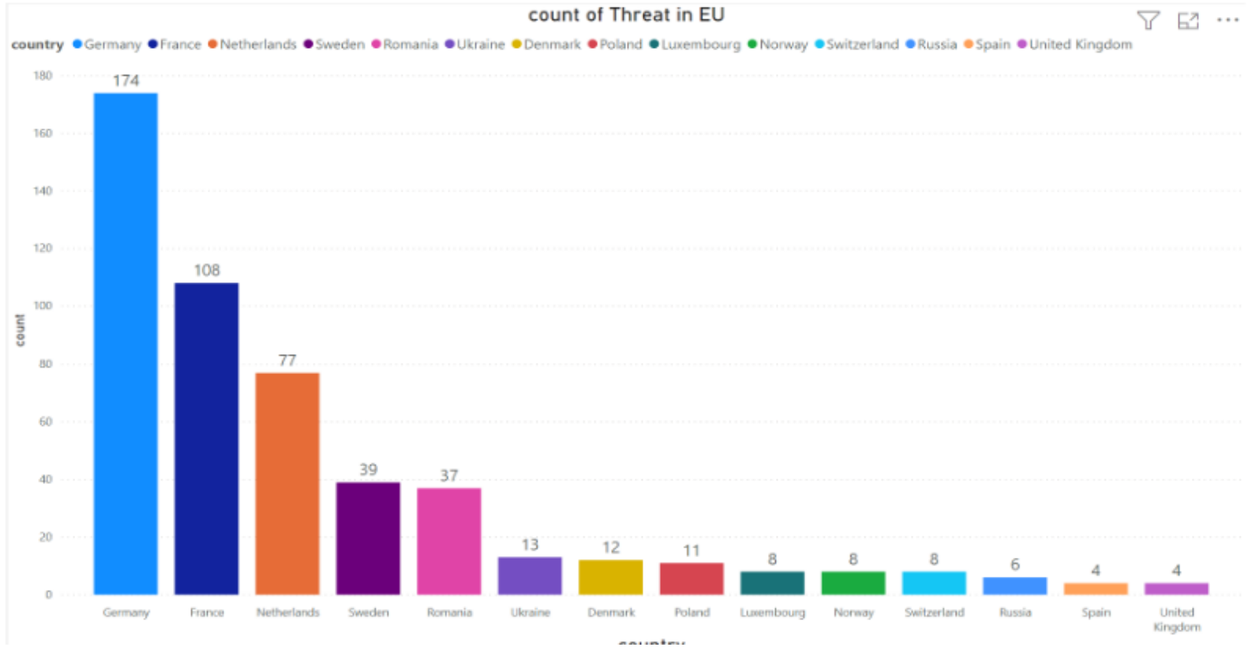


Figure 11: Shows actions and threat types by date







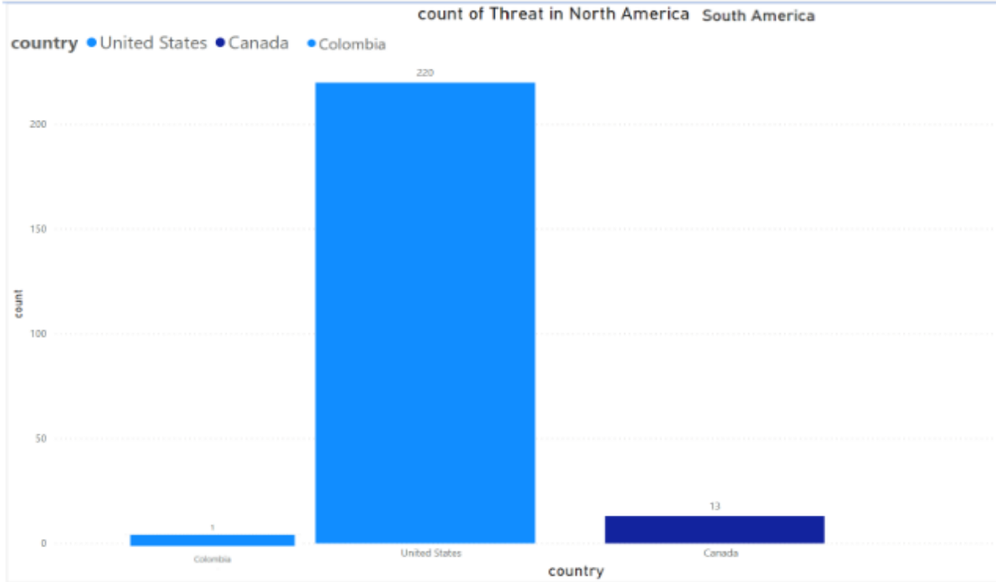


Figure 12: Shows threat's continents and countries

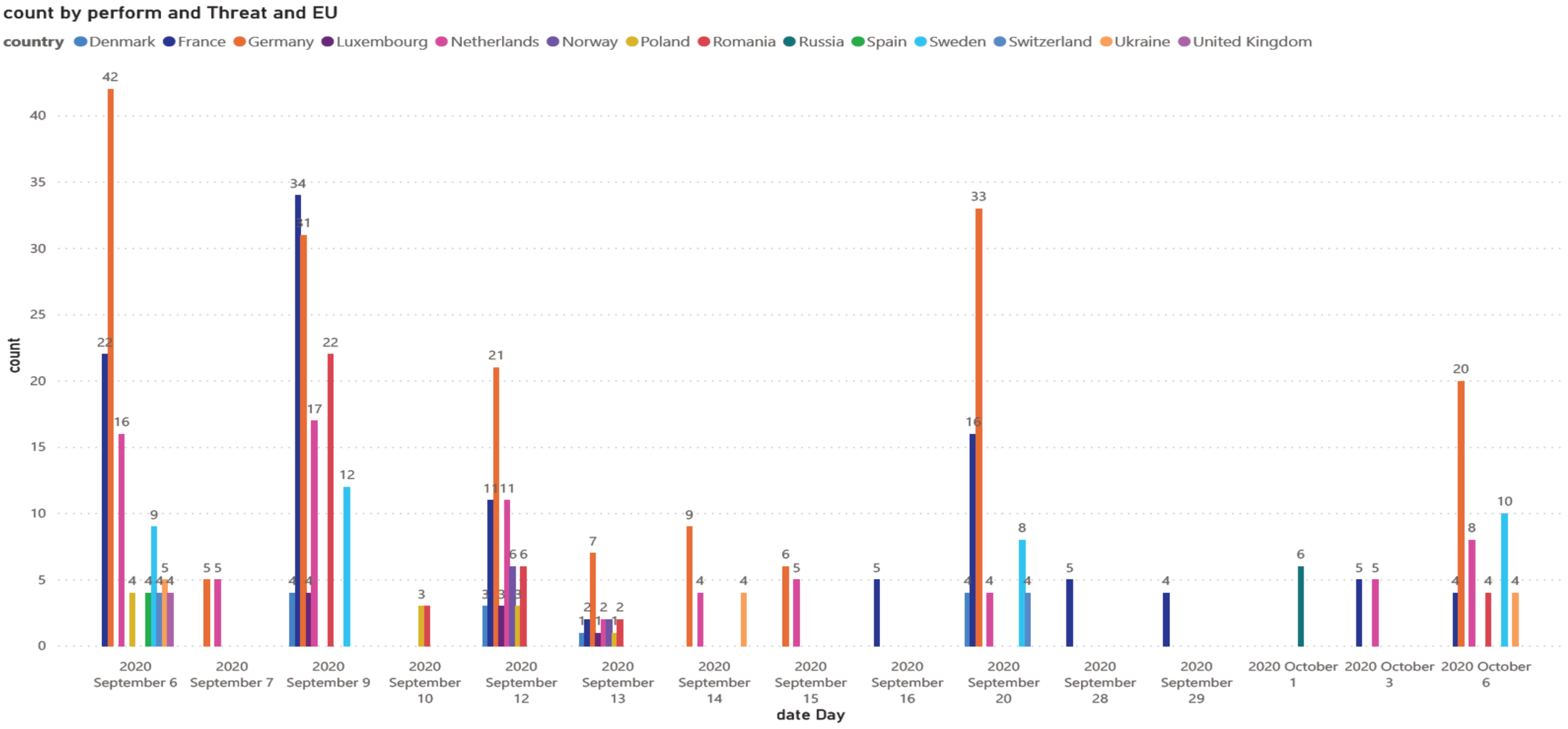
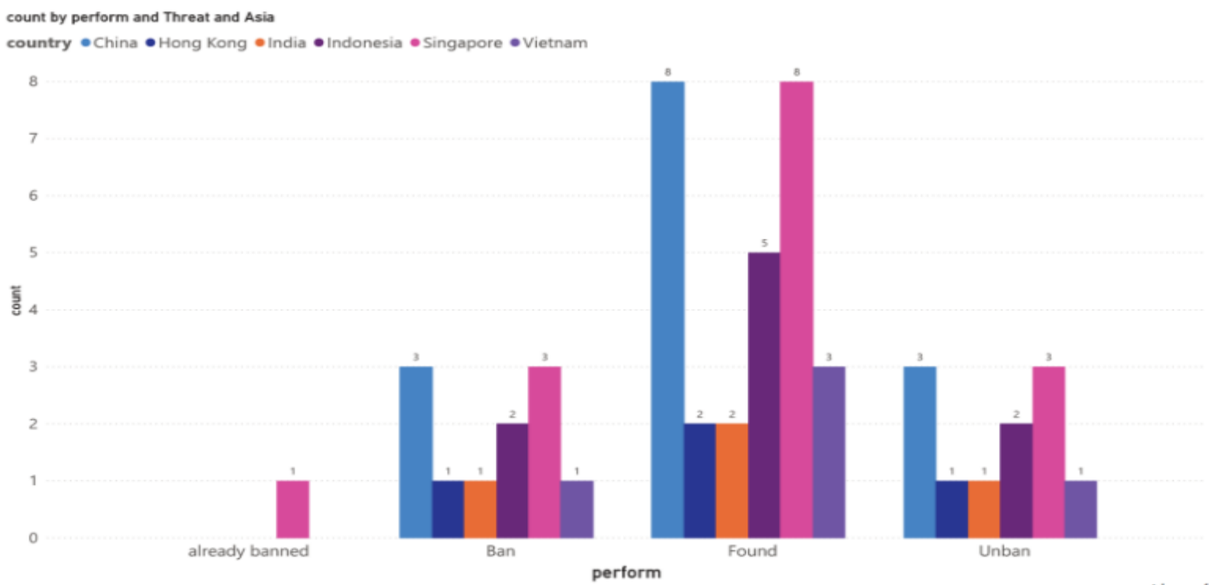
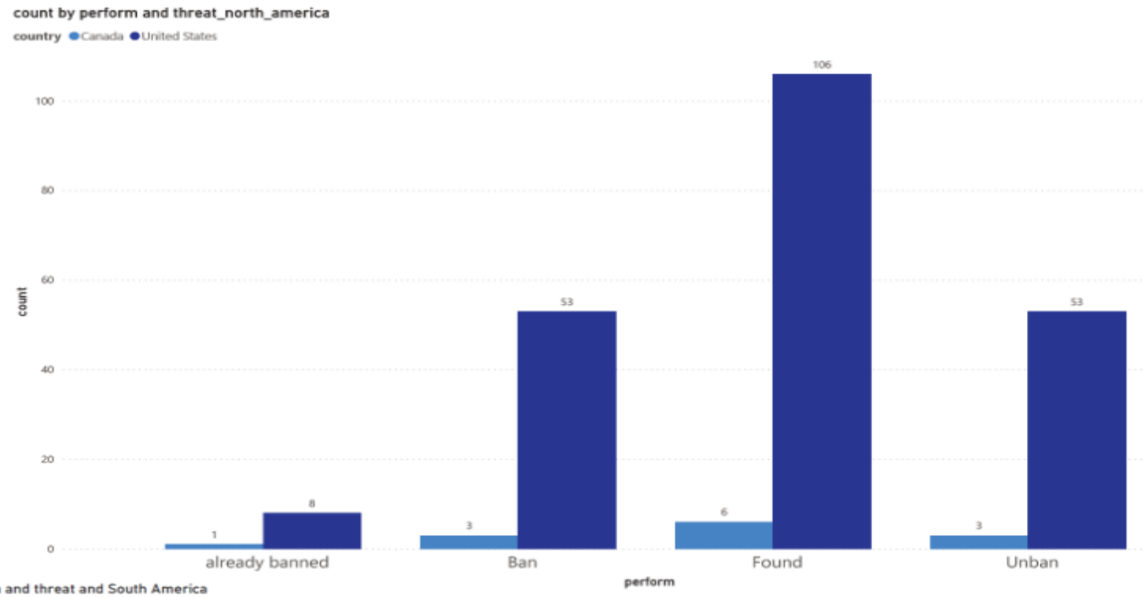


Figure 13: Shows threats from Europe continent by dates





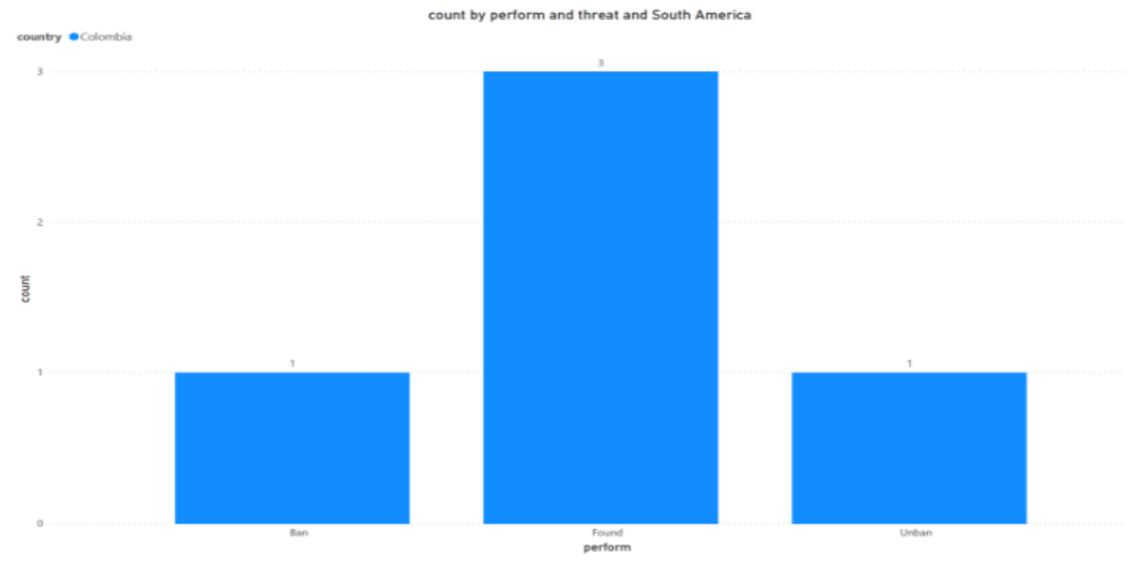


Figure 14: Shows threats from Asia, South and North America

There are totally 656 known attackers see figure[15] from 3 different continent. Europe has 435 known attackers from 11 different countries and Germany has 174 known attackers. North America has 217 known attackers from 2 different countries and USA has 204 known attackers. Asia has 1 known attacker from 1 country and India has 1 known attackers. See figure[16] for some more details.

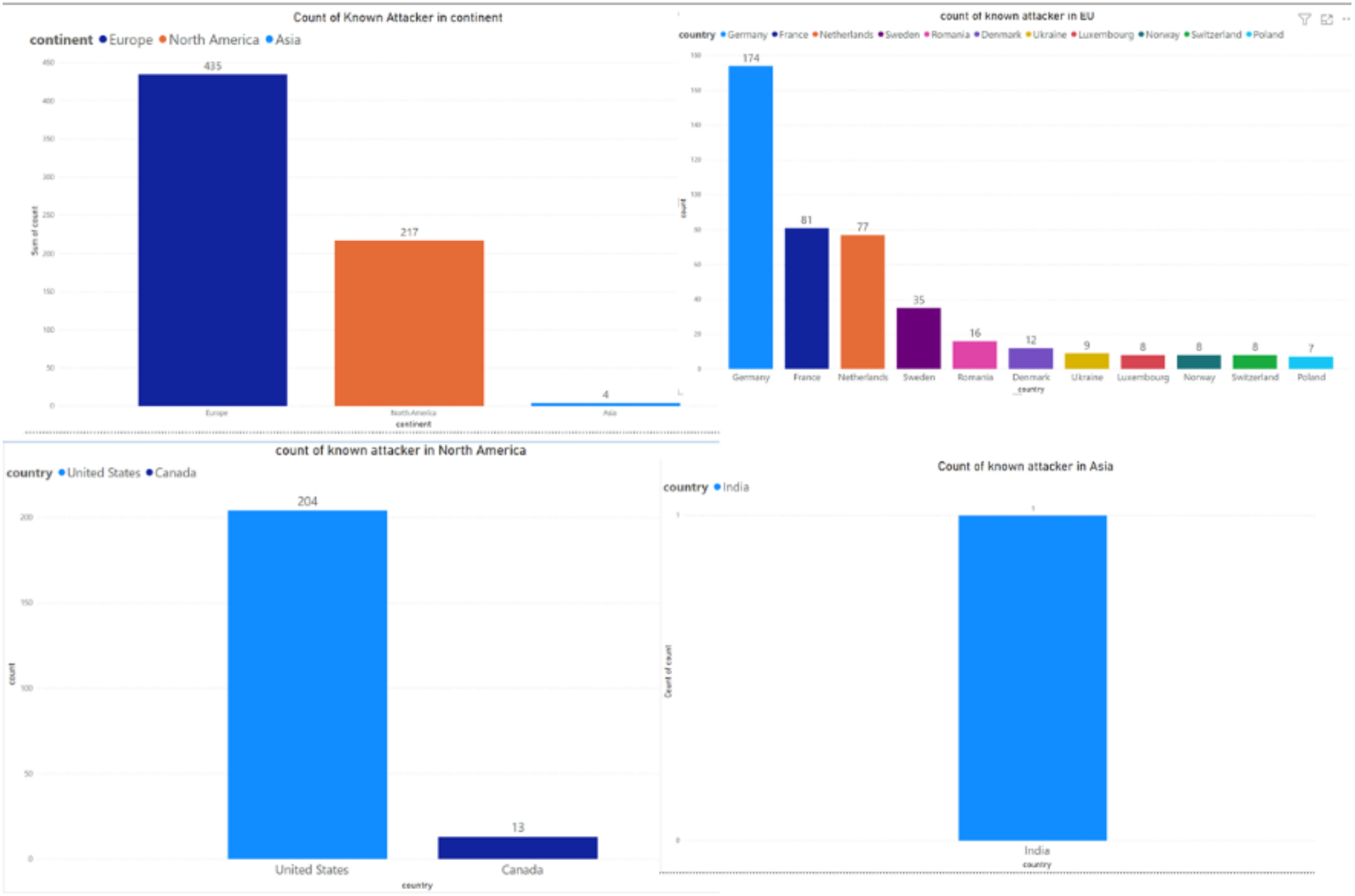
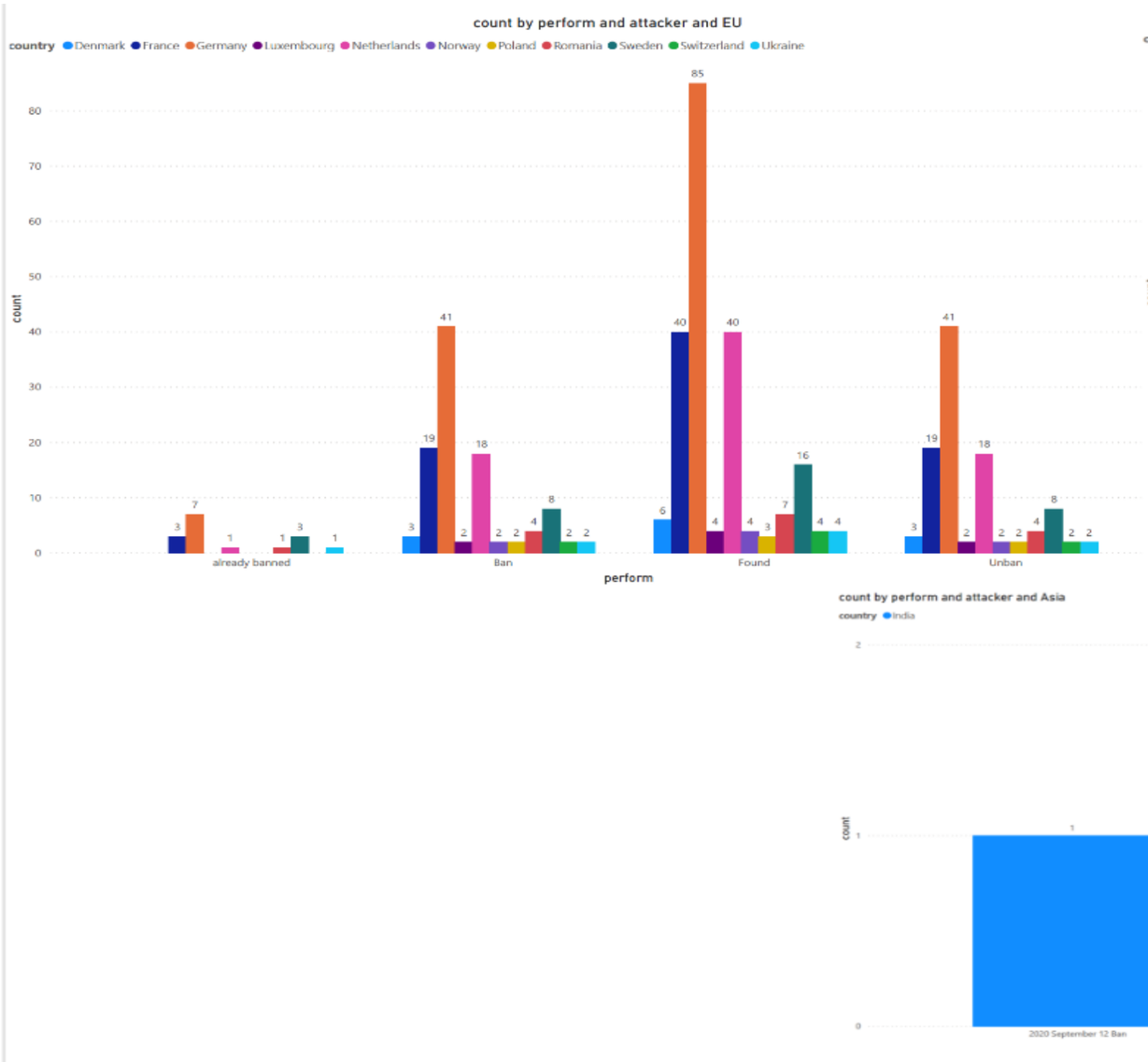
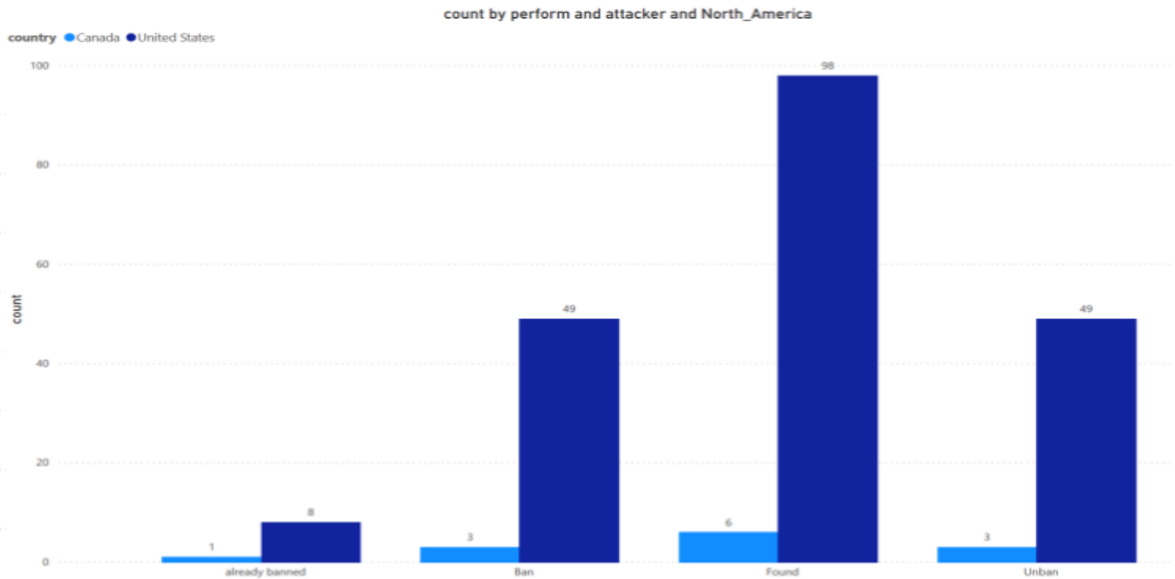


Figure 15: Shows attacker with country names





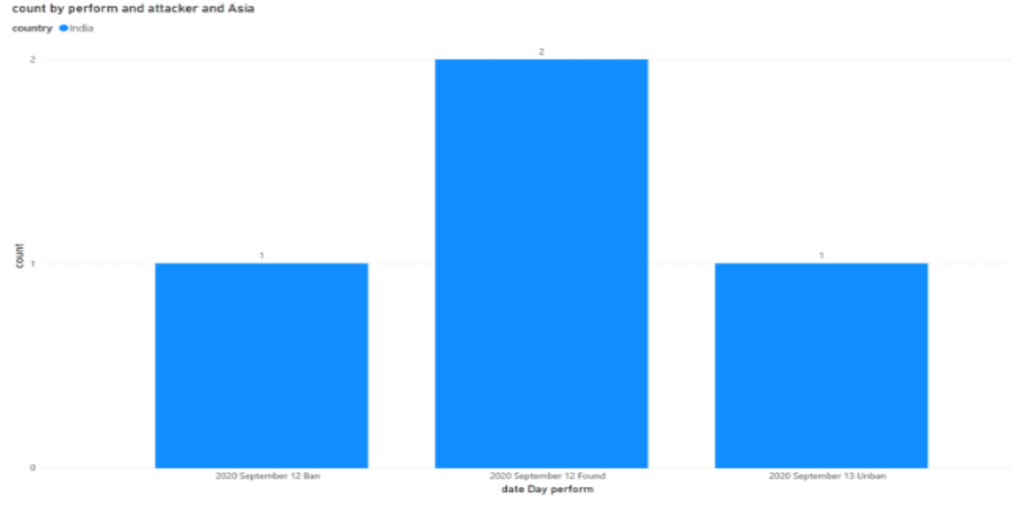


Figure 16: Shows attacker from Europe, North America and Asia, groupped by perform on the server-side

There are totally 484 known abusers from 3 different continents, see figure[17]. There are 235 known abusers in Europe. They are coming from UK. There are 230 known abusers in North American continent. They are coming from Panama and USA. There are 19 known abuser in Asia. For more details see figure[18]

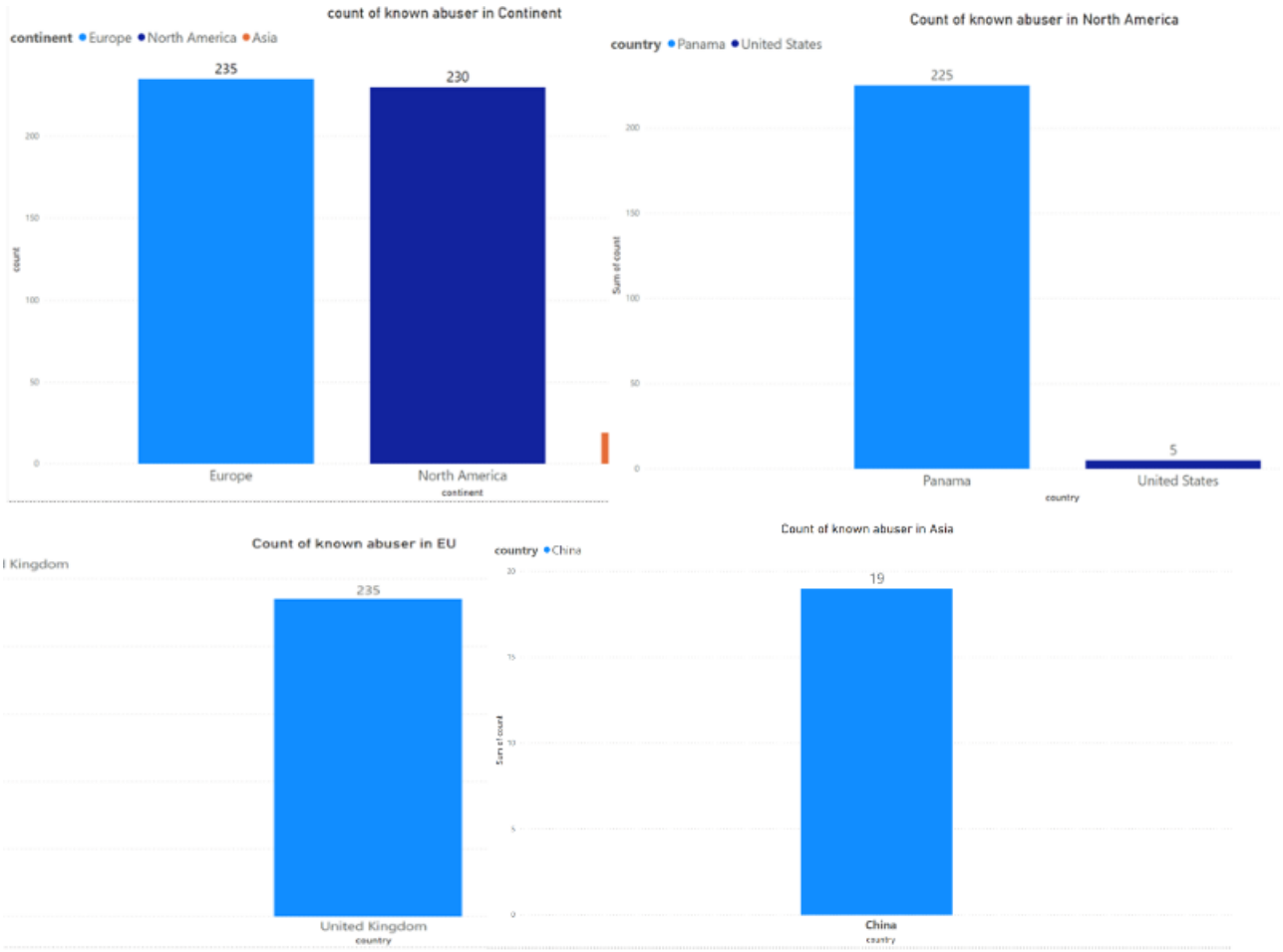
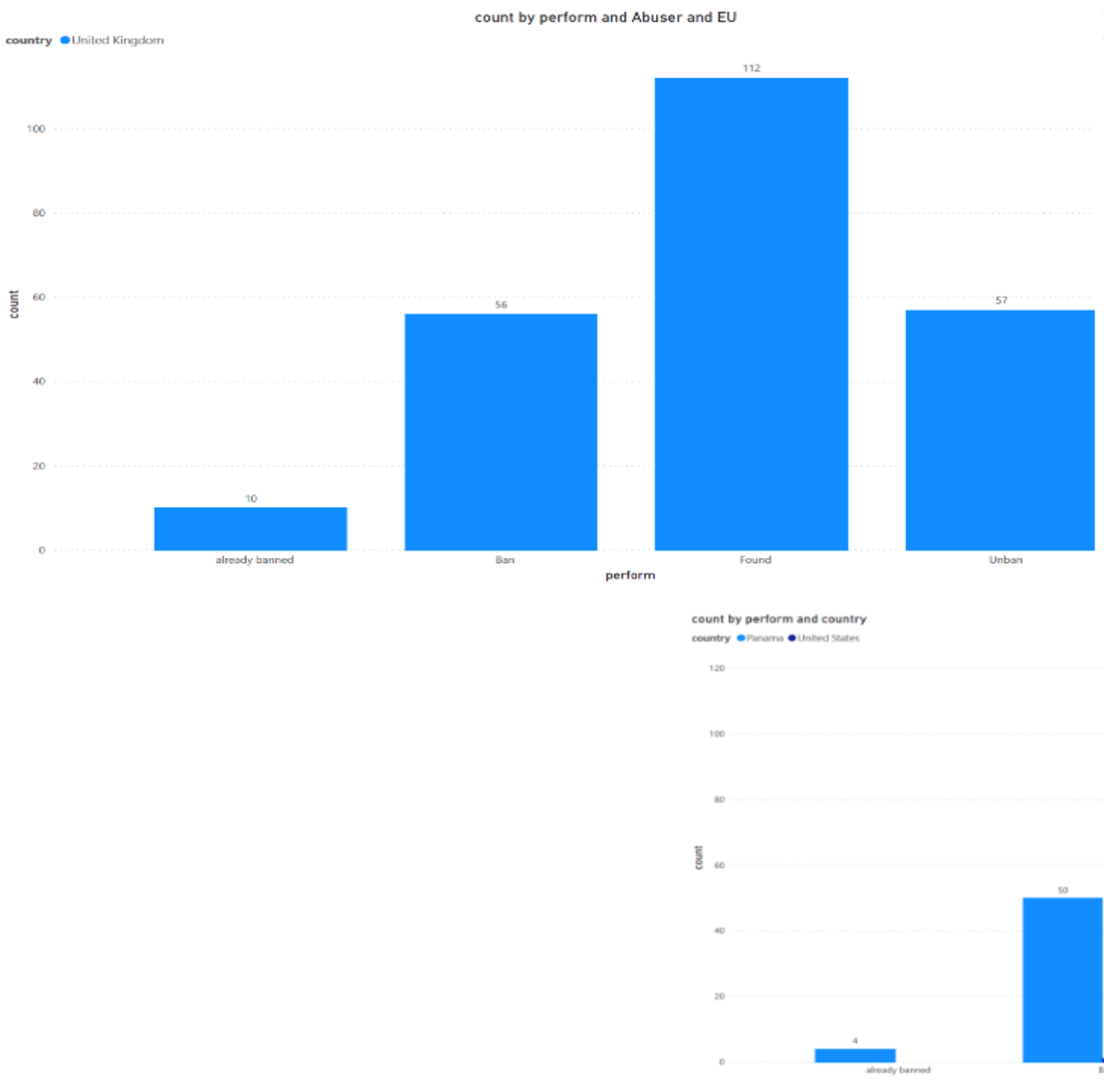
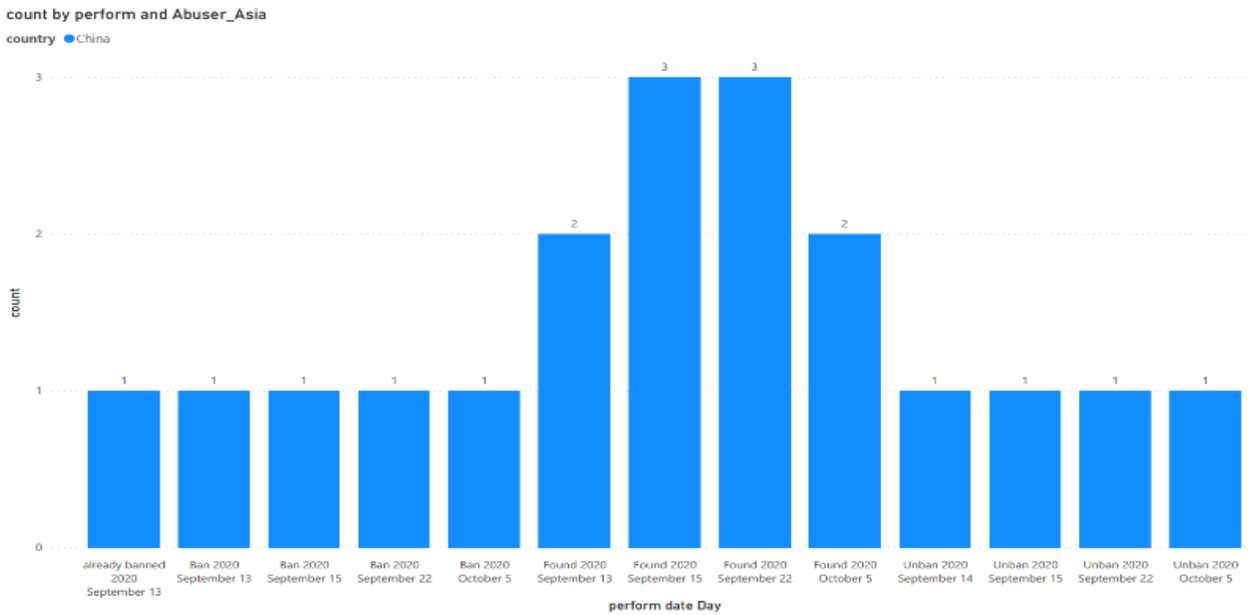


Figure 17: Shows abuser with country names





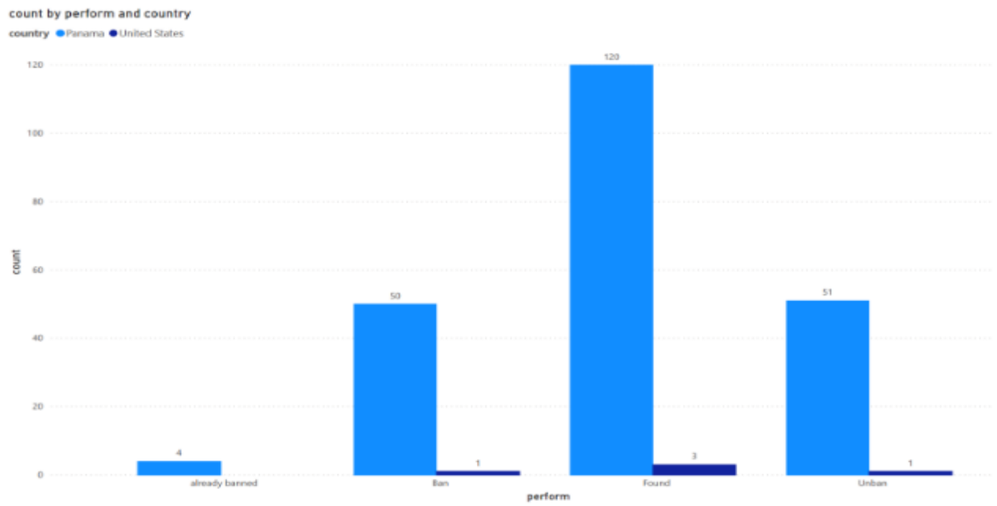


Figure 18: Shows abuser from Europe, Asia and North America

4.3 Bad IPs

After we removed duplicates users and figure[19] shows the users which is only threat. The x-axis shows the IPs and the y-axis shows the countries names where the user come from. We groupped the IPs by perform on the server-side, which some of the user is "Already banned" but they attempts to log in. On the country figure[20] describes the user who is only abuser with unique IPs. Last but not least figure[21] shows the user who is a known source of malicious, is threat and at the same time is tor network user. These IPs should be blocked to get access to the server becuase they are harmful. This three figures shows the users with unique IPs.

Among those 71824 visitors, we have found a total of 798 threats from 4 different continents, and the most of threat is from Europe. 509 threats are coming from 14 different European countries, and the most threat activities come from Germany. 51 threats are coming from 6 different Asian countries, and the most of threat activities are caming from Singapore. 233 threats are coming from 2 different North American countries and the country with the highest threat in the USA. 5 threats are coming from 1 South American country, and the country is Colombia.

The data we have includes not only threats but known attackers and known abusers as well. There are a total of 656 known attackers from 3 different continents. Europe has 435 known attackers from 11 different countries and Germany has 174 known attackers. North America has 217 known attackers from 2 different countries among them the USA has 204 known attackers. Asia has 1 known attacker and it is from India. Despite these, there are found 484 known abusers from 3 different continents. There are 235 known abusers in Europe and whiches are coming from the UK. There are 230 known abusers in the North American continent and whiches are coming from Panama and the USA. Finally, There is 1 known abuser in Asia and it is from China.

When we analyze these 3 dangerous activities, namely Threat, Known Attacker, Known Abusers, we can see that dangerous activities known as a threat and known attacker form a common set and this common set consists of 14 different states and 114 different IP addresses. By adding the activity called known abuser to these, we get a list of 24 different states and 152, see figure[22] different IP addresses when we analyze them. This list has been obtained after all the data analysis we have done from top to bottom and we believe that these IP addresses are dangerous for the Server. Figures of all the data we have discussed in this paragraph are available above.

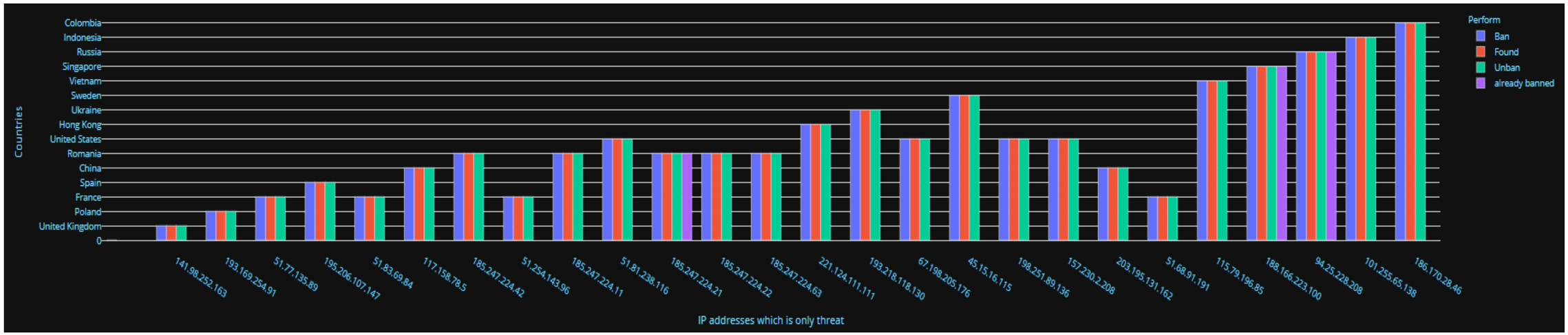


Figure 19: Shows IPs which is threat

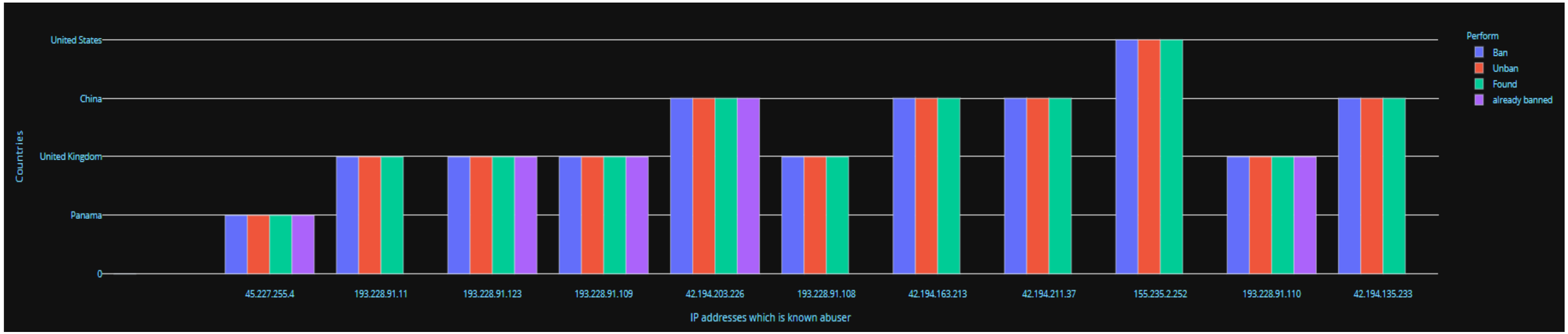


Figure 20: Shows IPs which is known abuser

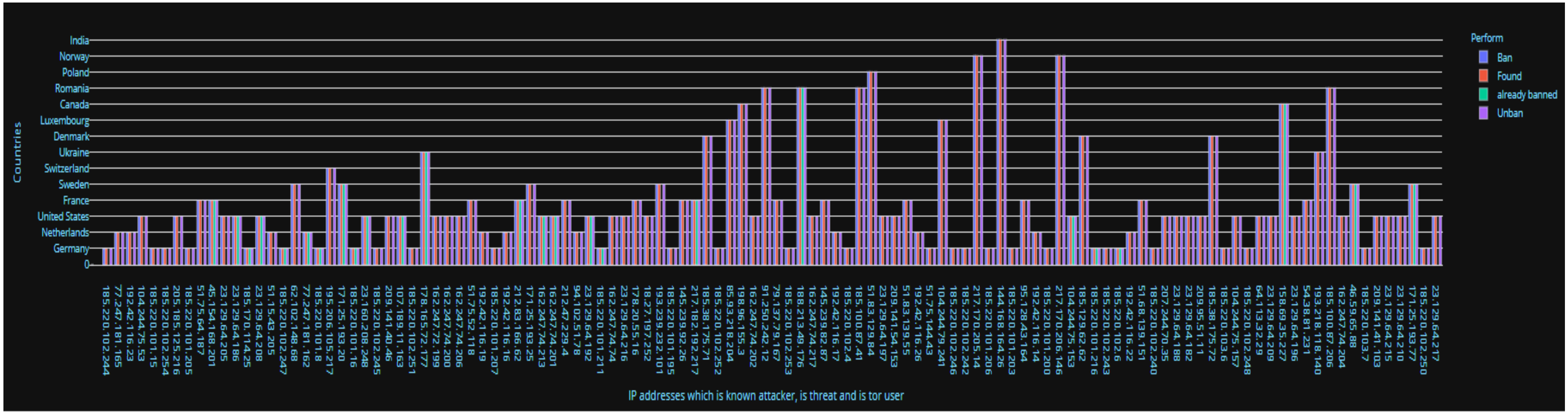


Figure 21: Shows the same IPs which is threat, is tor network user and is knwon attacker

Hong Kong	Spain	India	Switzerland	Panama	Colombia	Vietnam	Singapore	Russia	Indonesia
221.124.111.111	195.206.107.147	144.168.164.26	195.206.105.217	45.227.255.4	186.170.28.46	115.79.196.85	188.166.223.100	94.25.228.208	101.255.65.138

Poland	Luxembourg	Canada	Norway	Denmark	Ukraine	China	Romania	Sweden	United Kingdom
193.169.254.91	85.93.218.204	198.96.155.3	217.170.205.14	185.38.175.71	193.218.118.130	117.158.78.5	185.247.224.42	45.15.16.115	141.98.252.163
51.83.129.84	104.244.79.241	158.69.35.227	217.170.206.146	185.129.62.62	178.165.72.177	203.195.131.162	185.247.224.11	62.102.148.68	193.228.91.123
				185.38.175.72	193.218.118.140	42.194.203.226	185.247.224.21	171.25.193.20	193.228.91.11
						42.194.163.213	185.247.224.22	171.25.193.25	193.228.91.109
						42.194.211.37	185.247.224.63	193.239.232.101	193.228.91.108
						42.194.135.233	91.250.242.12	46.59.65.88	193.228.91.110
							188.213.49.176	171.25.193.77	
							185.100.87.41		
							185.100.87.206		

Netherlands	France		Germany		United States		
77.247.181.165	51.77.135.89	178.20.55.16	185.220.102.244	51.75.144.43	51.81.238.116	162.247.74.206	209.95.51.11
192.42.116.23	51.83.69.84	145.239.92.26	185.220.101.15	185.220.102.246	67.198.205.176	162.247.74.213	104.244.75.157
51.15.43.205	51.254.143.96	217.182.192.217	185.220.102.254	185.220.102.242	198.251.89.136	162.247.74.201	64.113.32.29
77.247.181.162	51.68.91.191	79.137.79.167	185.220.101.205	185.220.101.206	157.230.2.208	23.129.64.192	23.129.64.209
192.42.116.19	51.75.64.187	145.239.82.87	185.170.114.25	185.220.101.203	155.235.2.252	162.247.74.74	23.129.64.196
192.42.116.16	45.154.168.201	51.83.139.55	185.220.102.247	185.220.101.200	104.244.75.53	23.129.64.216	162.247.74.204
94.102.51.78	51.75.52.118	95.128.43.164	185.220.101.8	185.220.101.216	205.185.125.216	18.27.197.252	209.141.41.103
192.42.116.17	212.83.166.62	51.68.139.151	185.220.101.16	185.220.102.243	23.129.64.205	162.247.74.202	23.129.64.215
192.42.116.26	212.47.229.4	54.38.81.231	185.220.102.245	185.220.102.6	23.129.64.186	162.247.74.217	23.129.64.210
192.42.116.14			185.220.102.251	185.220.102.240	23.129.64.208	23.129.64.197	23.129.64.217
192.42.116.22			185.220.101.207	185.220.103.6	23.160.208.248	209.141.54.153	
			185.220.101.211	185.220.102.248	209.141.40.46	104.244.75.153	
			185.220.101.195	185.220.103.7	107.189.11.163	207.244.70.35	
			185.220.102.252	185.220.102.250	162.247.72.199	23.129.64.188	
			185.220.102.253	185.220.102.4	162.247.74.200	23.129.64.182	

Figure 22: Lits of bad IPs

4.4 Database

We have designed a relational database for our data, and with this, we can organize data in tables. The table contains rows and columns where a row is a tuple or a record, and a column is an attribute. Our main objective of creating a database was to eliminate data redundancy, make sure data integrity and accuracy. We have designed and customized our database to suit our application. The type of relationship we are using between tables is one-to-many. We applied normalization rules to study whether our database is optimal and structured correctly. We used DBeaver Community platform database to connect to our database in postgreSQL and figure [23] is ER-Diagram which is created by DBeaver tools.

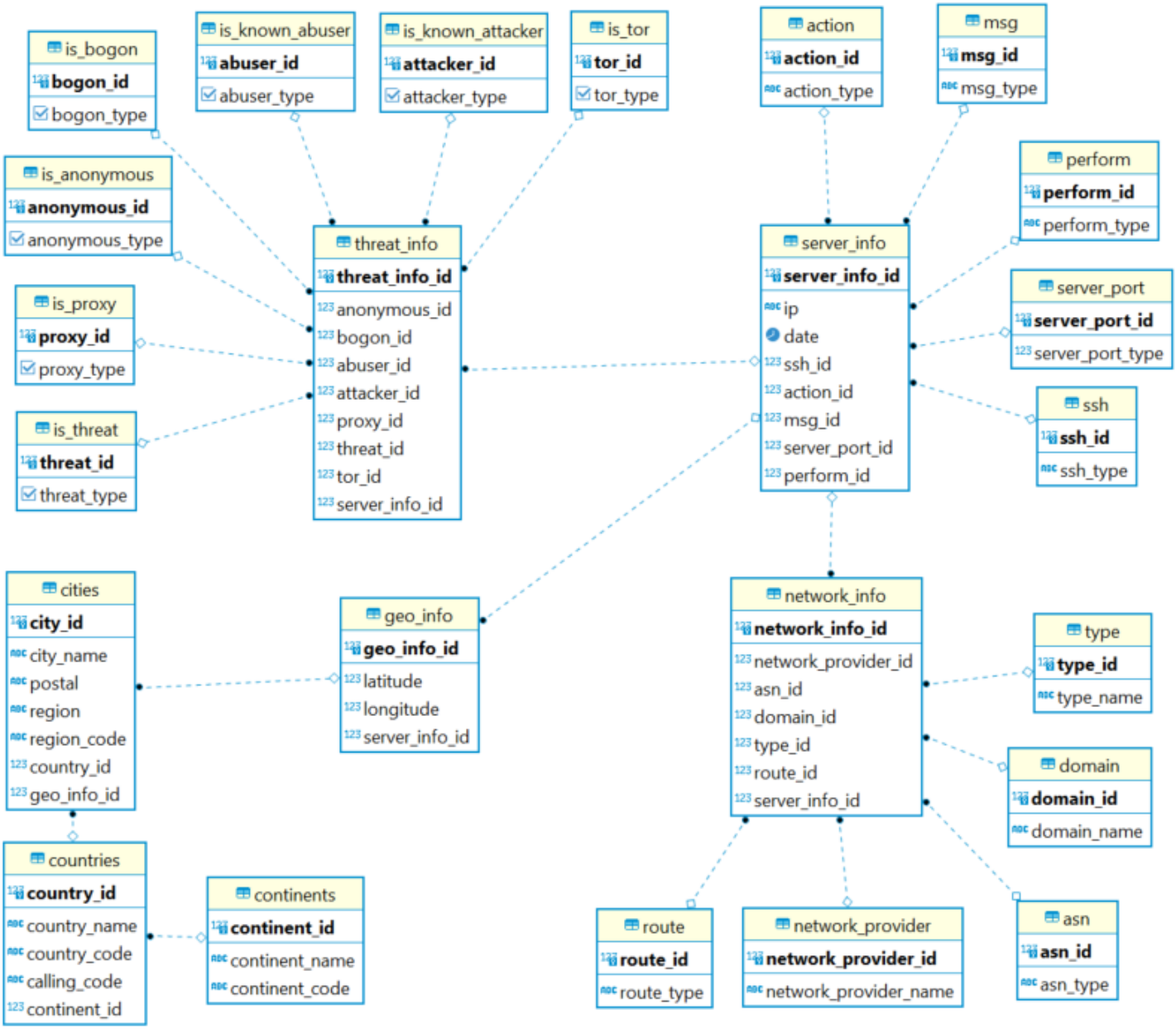


Figure 23: Database Schema

5 Conclusion

In conclusion, visitors in the categories known attacker, known abuser, and threat present a certain level of danger to our server. The security of the server is always of the utmost importance to us. For these reasons, we should always be careful with visitors in the above category. Apart from paying attention to countries where known attackers such as United States, Germany, France, Netherlands, Sweden, Romania, China are concentrated, it is beneficial to follow them carefully in countries such as the UK where there are many known abusers.

While doing this project, we think we haven't dealt with some issues deeply enough. As we mentioned, we have more than 71000 IP location information and information about so many cities and their locations. Therefore, we realized that it would not be possible in such a short time to analyze this information in more detail on a city basis. But in a future study, we believe that doing this in more detail will be a very useful factor in creating an ML, AI model.

When we think about a future study, we can create a model that can create a Machine Learning model based on all the information and analysis we have, and with this model, we can quickly categorize each visitor. Code for this project will be availabe on [MN].

References

[ipd] ipdata.co. *IP Geolocation and Proxy Detection API*. URL: `https://docs.ipdata.co/`. (accessed: October 25, 2020).

[Max] MaxMind.com. *GeoIP2: City and Country CSV Databases*. URL: `https://dev.maxmind.com/`. (accessed: October 25, 2020).

[MN] Rahmat Mozafari and Nimetullah Necmettin. *Data Management: Data Analysis and Visualization*. URL: `https://github.com/Mozafari1/data-management`. (Is availabe from: November 2, 2020).