**Executive summary**

**Finding**

I just found out that all the log data are recorded with the IP address, access date and time, HTTP codes, user agent information, method ,etc. It looks a bit boring and tiresome by just looking at this essentially a text file. And it was hard to use text file for analysis. It is not a typical tabular form like CSV file and that makes it hard to see what information is even there. Even after you managed to import this text data into a tabular form, it still doesn’t seem to give you much interesting information.However, there are some useful information burried inside this log data if you look deeper. For example, you can technically get location information out of the IP addresses because each unique IP address is assigned to each client machine (PC) based on its geographical location. And if you know how to dissect the URLs you can not only see where the requests are referred or originated from and understand the users’ behavior based on the sequences of the requests. Then I had to change a text file in to csv file.

**Data description**

The data contains 3 datasets collected from web vulnerability scanners, for example, Acunetix, Netsparker and w3af and i combined all data sets in to a xss\_sqli\_attacks\_detection .the data set has the following fields

ip\_address The IP address is a 32-bit number that uniquely identifies a network interface on a machine. An IP address is typically written in decimal digits, formatted as four 8-bit fields separated by periods. Each 8-bit field represents a byte of the IP address. This form of representing the bytes of an IP address is often referred to as the **dotted-decimal format**. The bytes of the IP address are further classified into two parts: the network part and the host part.the data type is IPV4

timestamp data type is used for values that contain both date and time parts.

Method.it consists of symbolic data

requested\_url indicates a file name ( index.php), a component, and search. its consist of symbolic data.

status\_code it consists of numeric data

bytes\_transferred. it consists of numeric data

referrer\_url . indicates a protocol ( http ), a IP ( 192.168.4.161 ), and a file name ( index. php).

protocol. it consists of symbolic data

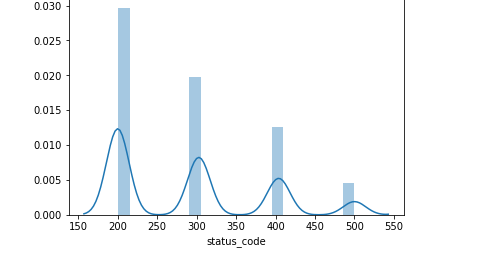
|  |  |
| --- | --- |
| **Web Vulnerability Scanner** | Data size |
| netsparker | 7314 |
| acunetix | 6539 |
| w3af | 3996 |
| xss\_sqli\_attacks\_detection.csv | 17849 |

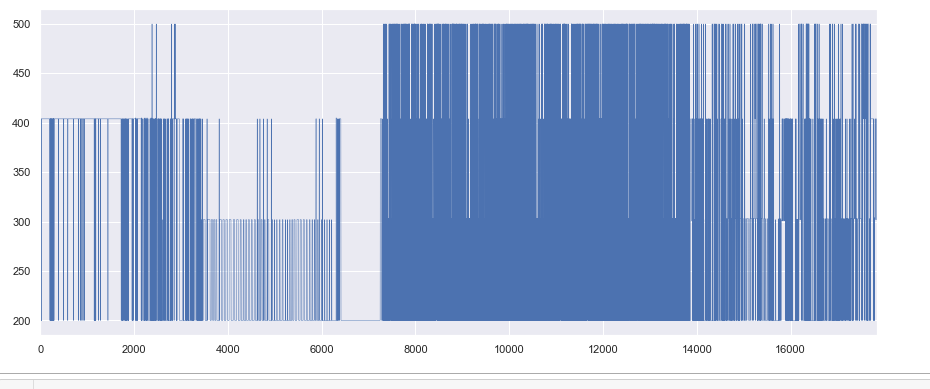
**data analysis and visualisaton.**

from status\_code distribution

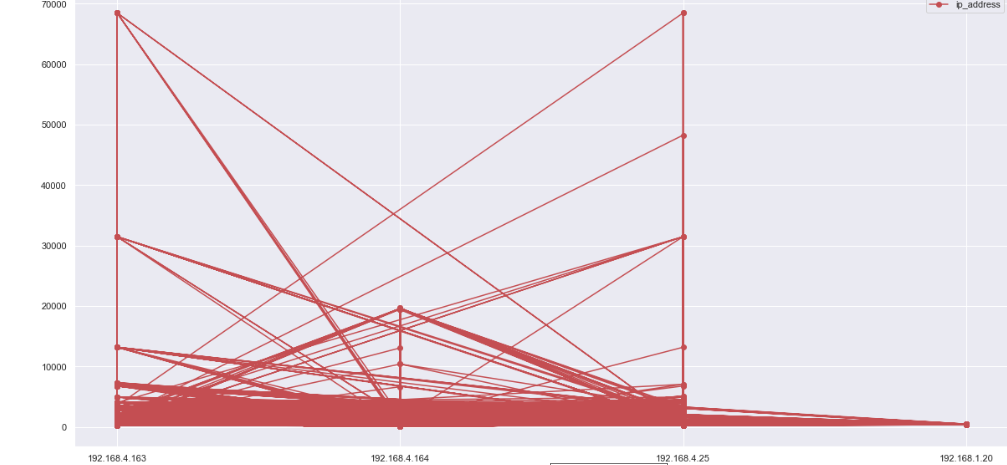
Its visually seen that the 200 error message has the highest number of requests meaning that most of the requests submitted where successfully replied without any attack.

The 404 massage comes next with the second highest replies which indicates that the request was unsuccessful due client error side due to attack between the user and the web servers indicating some kind of vulnerabilities in the web browsers like traffic, denial of service attacks etc.

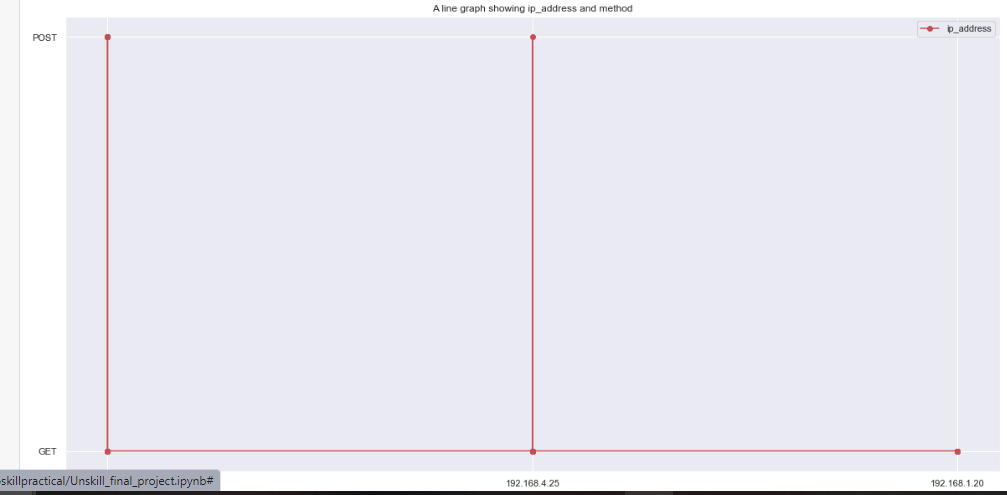




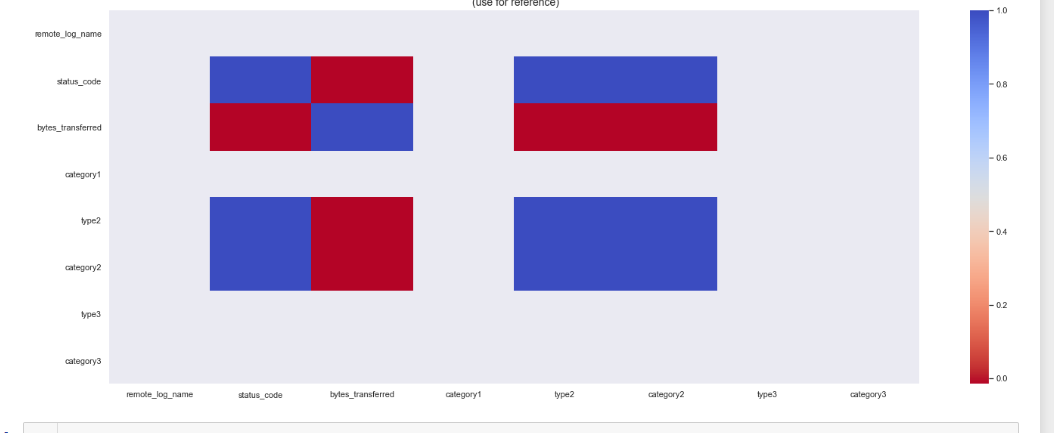
From the graph above, it indicates that the new csv file has a lot of data mostly concerned in the 8000 and 14000 and from 0 to 8000 there is less concentration of data



From the graph above ,the data set had very many column hence making it hard to clearly study the graph







It summarizes a large amount of data where the goal is to see patterns. In our example above, the observable pattern is that all the variables highly correlate with each other.

**Conclusion**

More rules are needed to identify vulnerability scans, SQLI injection, access log files for detections, XSS and SQLI attacks to enable proper manipulation of data.

According to type1 user detection, a rule-based model was not good enough to detect cross-site scripting and SQL injection

Datasets provided where large and could facilitate proper machine learning results.