

HSLU: LUCERNE UNIVERSITY OF APPLIED SCIENCES AND ARTS

DVIZ Main Project

Kevin Häusler

BSC Artificial Intelligence and Machine Learning

Tutor: Prof. Elena Nazarenko, PhD

January 1, 2025

Table of Contents

1	Work Table	2
2	About the Project	3
2.1	What is the project?	3
2.2	Data	3
2.3	Tools	3
2.3.1	Shodan	4
2.3.2	Streamlit	4
2.3.3	LaTeX	4
3	Motivation	5
3.1	Cyber Security	5
3.1.1	Homelab	5
4	Research	6
4.1	A Study on Internet of Things Devices Vulnerabilities using Shodan	6
4.2	Guides on how to use Shodan	6
4.3	Using Shodan to hunt down Ransomware Groups	6
5	Data	7
5.1	Getting the Data	7
5.2	Preparing the Data	8
5.2.1	Unexpected Issue	9
5.3	Analyzing the Data	10
5.3.1	What are the most commonly used Software?	11
5.3.2	What are the most commonly used ports?	12
5.3.3	Fake Security Attempts	13
5.3.4	Are there any weird/interesting findings that stand out?	13
5.4	Conclusions	14
5.4.1	What did I learn?	14

1 Work Table

Due to me (Kevin Häusler) being the sole member of this project I have omitted the "done by" column.

Date	Hours	Task description
22.10.2024	1	Downloading 9k Dataset from Shodan
07.11.2024	1	Setup the project, convert the json data to xlsx, setup latex environment
07.11.2024	1	Configure the initial streamlit setup
08.11.2024	1	Start writing the report + setup Github
15.11.2024	2	Research about Shodan
23.11.2024	1	Writing more functions in the code
08.12.2024	1	Downloading new 32k Dataset from Shodan
08.12.2024	2	Analysis of the new 32k Dataset
09.12.2024	4	Code Refactor and Data Analysis
10.12.2024	1	Update Report, start working on Hosting of Project

2 About the Project

2.1 What is the project?

The project is something akin to a cyber security analytic about the devices listed on shodan.io based on their location being Rotkreuz. It is supposed to show insights about what is accessible and possibly insecure.

There are also dynamic parts where you can filter the data.

2.2 Data

The data is generated from shodan.io. I do have a lifetime membership that grants me enough credits to request and download 10'000 entries. I did filter the data to be only from Rotkreuz which resulted in a little bit over 9000 datapoints that I am using for this project.

Why I am not using the entire 32'000 dataset is explained in my report where I go indepth about how I got the data.

2.3 Tools

Here is a list of the tools I am using:

PyCharm
Python 3.10
UV
Streamlit
Shodan CLI
Docker
Github
LaTeX

I have setup the project in pycharm with UV because I wanted to try it out, there is also a docker file in case you have docker installed to easily run this project without having to create your own environment. The requirements are only a few packages so they can also just be installed with the requirements.txt file.

For the main library I have decided to use Streamlit, I have tried it out in the past but never for a real project so I wanted to get more indepth experience with it. It has a great documentation and it lets me easily create dataframes and charts for this project. It also needs the Altair and openpyxl packages to work.

The whole report is written in the PyCharm IDE with LaTeX and uploaded to the Github repository. It is written in the report.tex file and compiled to report.pdf.

This setup is very useful in case I need to work from different machines. Writing the code and the report in the same environment is something I really like.

2.3.1 Shodan

Shodan.io or more commonly referred as just Shodan is a search engine tool used to provide a more comprehensive view of the internet, more specifically it crawls, scans and collects information from a wide range of devices that are public to the internet.

This includes servers, IOT devices, webcams, routers even industrial control systems.

This data can be browsed and used to "help" assess cybersecurity risks. It does not guarantee that your systems are secure or insecure, so do not mistake it as being a complete cybersecurity assessment.

I was lucky enough to get a "lifetime" membership for 5 USD, this gives me access to multiple tools and 100 query credits.

1 Query credit corresponds to 100 results which is why my initial dataset to try out the feasibility of this project only contains 9000 datapoints. For this project I decided to sign up for a freelancer membership so I get 10'000 query credits which I can use to get the entire dataset of Rotkreuz and maybe implement an interactive tool as well with the Shodan API.

2.3.2 Streamlit

This is not my first time trying out streamlit but it is my first time actually trying to create a complete project with streamlit.

I like that it is relatively easy to get started with and how good their documentation is. I have not done any special formatting for the charts and datatables though because the way my data is displayed does not require it.

2.3.3 LaTeX

I have been using Linux for quite a while so I do not have easy access to the office suite anymore. I also have used LaTeX in the past to create my math cheatsheets/documentation and I really like the way it works. A big plus is that I can set it up to work in my IDE which also lets me store everything in the same Repository as my code.

3 Motivation

The main motivation of this project was to use shodan.io and streamlit to create something interesting. I have rarely used my shodan membership which felt bad even though I only paid 5 USD for a lifetime membership.

Streamlit has been something I have been trying out for a personal project but I never really got into it as much as I would have liked due to other commitments.

3.1 Cyber Security

Another relevant motivation would be cyber security. This semester I am taking the Introduction to Cyber Security class which has been surprisingly fun and interesting. While this project does not equal to a real cyber security assessment I do want to try my best to create something that is informative and useful in that regard.

3.1.1 Homelab

The project itself is also motivated by my own homelab. I have multiple physical and virtual servers at home that run many different services that are private or open to the public. This is why I am also interested to see how others fare in this aspect.



Figure 1: Messy Picture of part of my homelab

4 Research

4.1 A Study on Internet of Things Devices Vulnerabilities using Shodan

There is a short research paper about the vulnerabilities of IOT devices where the author used shodan as a search engine to get the required data for his research.

4.2 Guides on how to use Shodan

There are many interesting and informative articles on medium.com and cyber security blogs about how to leverage Shodan for cyber security or more "nefarious" purposes like gaining access to unsecured webcams and NAS.

4.3 Using Shodan to hunt down Ransomware Groups

One of my favorite cybersecurity researcher and youtuber John Hammond is affiliated with the cybersecurity firm Huntress, which is how I came upon a blog post about using Shodan to hunt down Ransomware groups.

While it is not entirely relevant to my project it has been a very interesting read and I can recommend it for people that are interested in this.

<https://www.huntress.com/blog/using-shodan-images-to-hunt-down-ransomware-groups>

5 Data

5.1 Getting the Data

As mentioned I will be using Shodan to get the data about all the devices that are scanned in Rotkreuz. This is done with their search engine tool on <https://www.shodan.io/search?query=country>

The screenshot shows the Shodan search results page for the query "country:CH city:rotkreuz". The page displays 32,147 total results. On the left, there are filters for Top Ports, Top Organizations, Top Products, and Top Operating Systems. The main content area displays a list of search results, including IP addresses, hostnames, and various services like SSH, SSL Certificates, and HTTP. Each result includes a brief description and a date.

Figure 2: Shodan Search Results for Rotkreuz CH

Here we can see 32'147 results which I downloaded.

The screenshot shows the Shodan "Download Results" modal window. The modal displays the search query "country:CH city:rotkreuz" and the number of results "32065". It includes a "DOWNLOAD" button and a note that downloads may take several hours to complete. There is also an FAQ section on the right side of the modal.

Figure 3: Downloading the search results

5.2 Preparing the Data

The download returns a *.json.gz archive which I have renamed to shodan_data_unprepped.json.gz for this project.

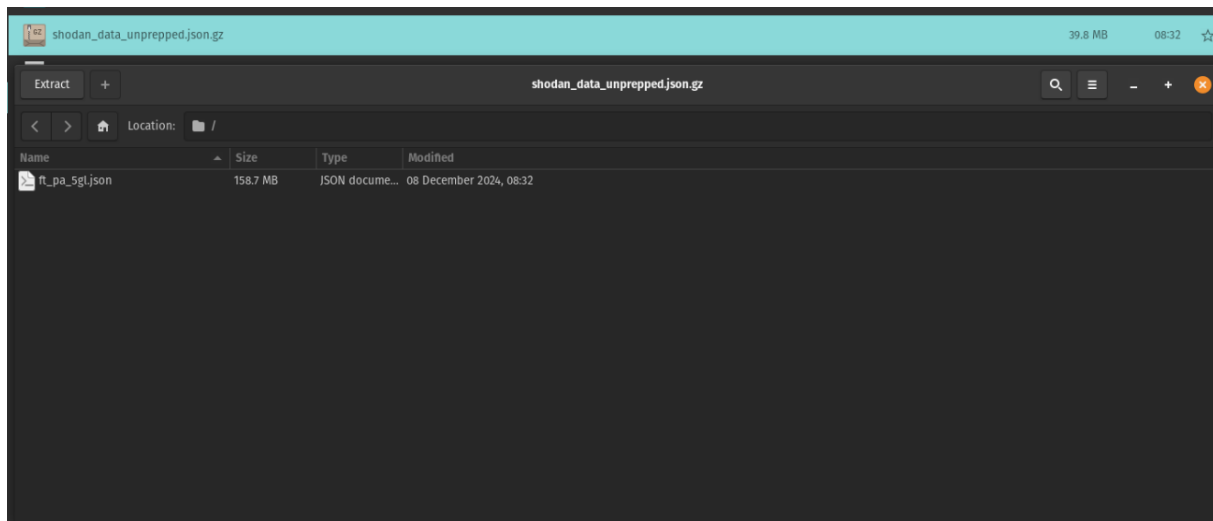
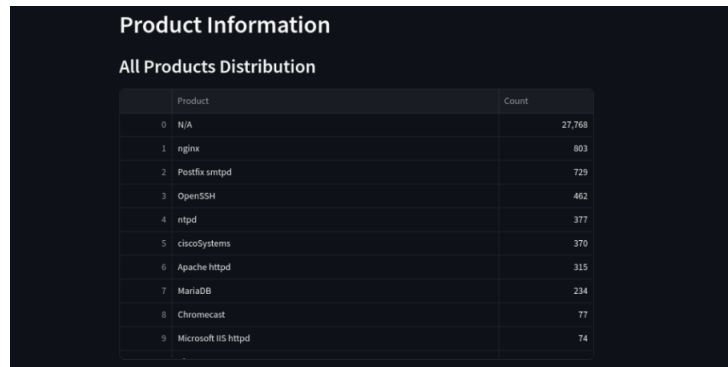


Figure 4: The Downloaded Archive
<https://www.shodan.io/search?query=country>

We can convert this with the following command: "convert shodan_data_unprepped.json.gz xlsx to an xlsx file which we rename to shodan_data_32k.xlsx.

5.2.1 Unexpected Issue

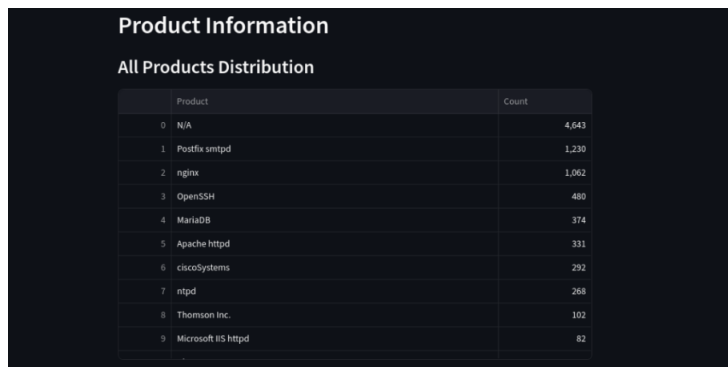
But what is this? We suddenly have a lot (27'768) of N/A entries in our dataset.



All Products Distribution	
Product	Count
0 N/A	27,768
1 nginx	803
2 Postfix smtpd	729
3 OpenSSH	462
4 ntpd	377
5 ciscoSystems	370
6 Apache httpd	315
7 MariaDB	234
8 Chromecast	77
9 Microsoft IIS httpd	74

Figure 5: 27'000 N/A with the 32k Dataset

When I compare this with the 9k Dataset that I downloaded on 22.10.2024 we only have 4'643 N/A entries.



All Products Distribution	
Product	Count
0 N/A	4,643
1 Postfix smtpd	1,230
2 nginx	1,062
3 OpenSSH	480
4 MariaDB	374
5 Apache httpd	331
6 ciscoSystems	292
7 ntpd	268
8 Thomson Inc.	102
9 Microsoft IIS httpd	82

Figure 6: Only 4'643 N/A with the 9k Dataset

An indepth look at the Shodan search engine shows an organisation called Digital Assets AG with 23'860 datapoints, and while they have a lot of open ports they do not return any data.

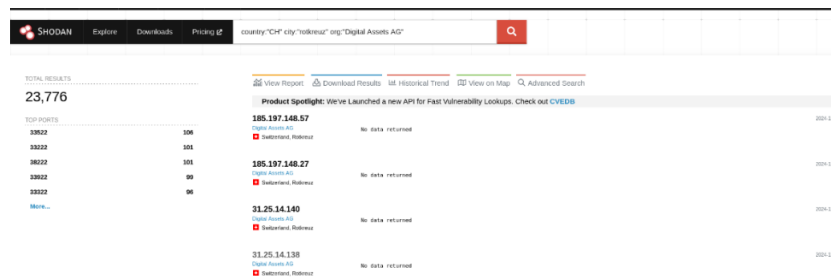


Figure 7: Digital Assets AG Rotkreuz CH Search

An overview of one of these IP can be looked at here <https://www.shodan.io/host/185.197.148.27>

When we look at the company itself we can see that they have 77'280 servers in Germany

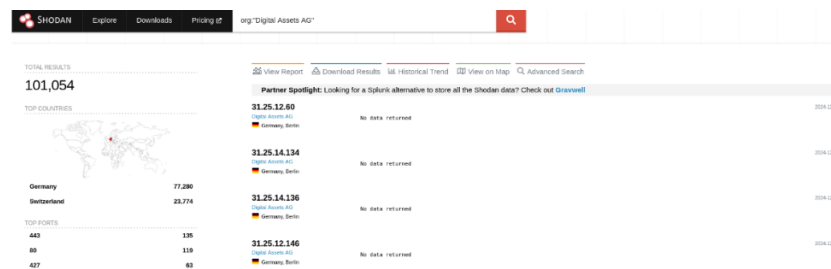


Figure 8: Shodan Digital Assets AG

s The ASN lookup confirms that it is part of the Google Cloud Platform. While I do not have any confirmation it is highly likely that they are renting or colocating servers (in this case VPS) in the new CKW Datacenter in Rotkreuz.

To be safe I will use the 9k dataset from 22.10.2024 for this project. Because if I were to use the 32k Dataset I would have way too many empty datapoints that I cant write about.

5.3 Analyzing the Data

Now that we have our data we can start analyzing it. First I want to check some general information like:

- What are the most commonly used Software?
- What are the most commonly ports?
- Are there any weird/interesting findings that stand out?

5.3.1 What are the most commonly used Software?

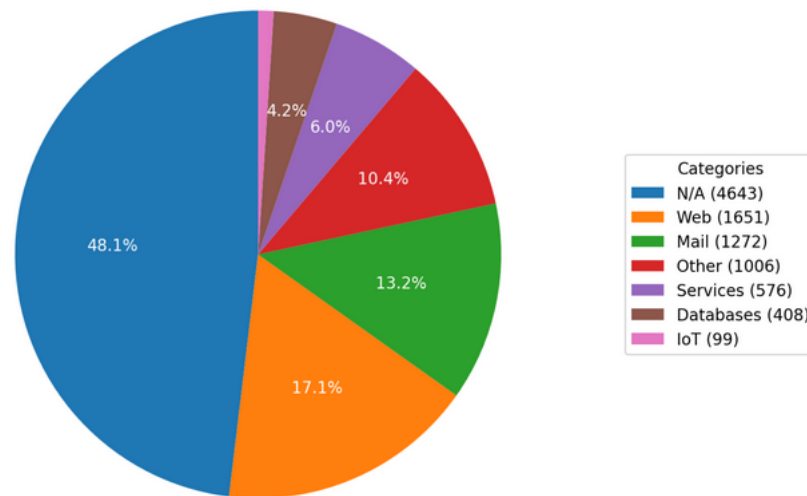


Figure 9: Product Pie Chart

A simple Pie Chart that was created by grouping regex searches like web/nginx/apache/httpd for web and other keywords for other categories lets us see which Products are commonly used. From this Pie Chart we can see that eMail and Webservices are used the most, this makes a lot of sense because they have to be publicly accessible to function correctly.

We can also see various Services ("dns", "snmp", "ssh", "remote desktop protocol", "rdp") and databases being publicly accessible. Depending on the service and the way it has been made accessible this can be a serious security risk. At our company we do offer external RDP but only via a remote desktop gateway with ssl termination. This is because raw RDP connections are very insecure and can be exploited.

5.3.2 What are the most commonly used ports?

A closer look at the Port distribution shows that the main use is for web- (port 443 and 80) and email services (port 25). There are some concerning ports like DNS (53) that are not recommended to be accessible publicly in most cases.

This open DNS port can be exploited by using the exposed DNS servers to DDoS other systems by sending tons of fake DNS responses in a DNS amplification attack.

Thankfully most of the exposed DNS ports are from web hosting companies and datacenters that should know the risks and how to prevent them. There are still 19 open DNS Ports from WWZ ISPs that could be residential that are worrisome.

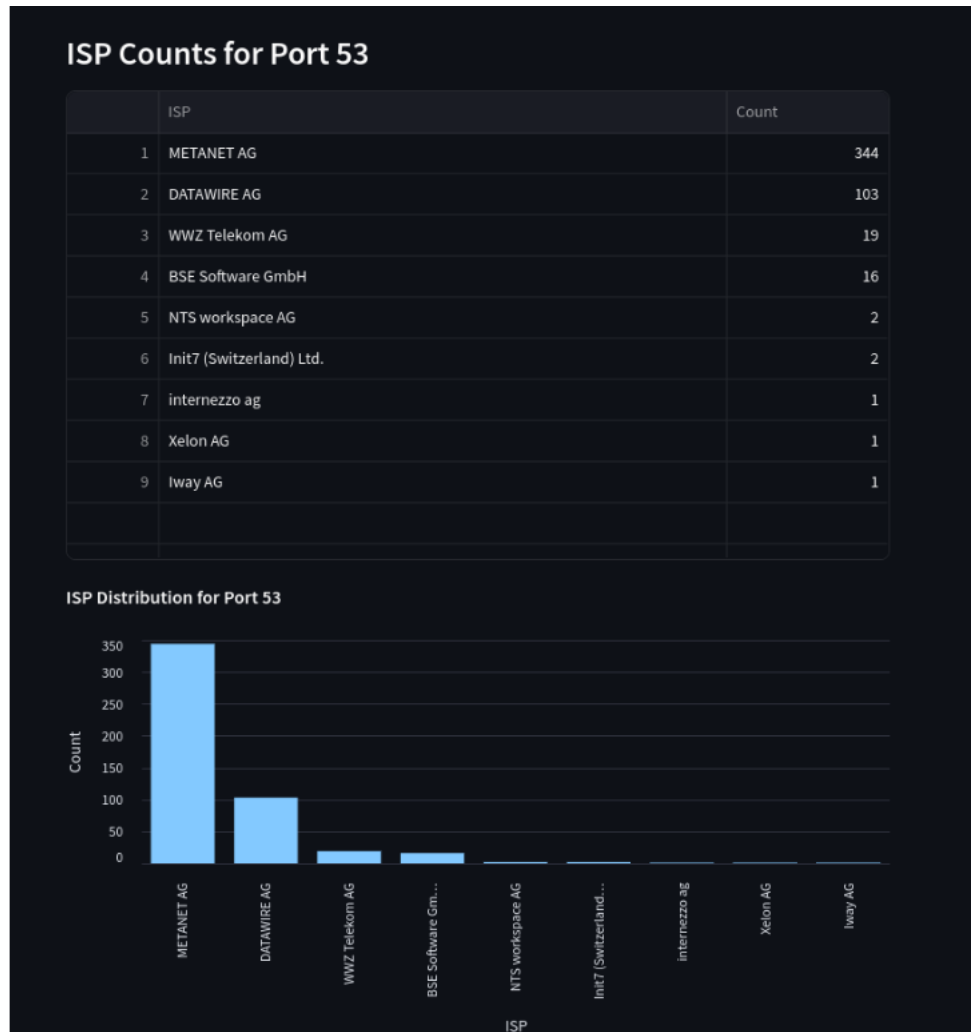


Figure 10: Port 53 Distribution by ISP

5.4 Conclusions

5.4.1 What did I learn?

As someone that has a huge homelab with services that I try to secure as best as possible I learned that in comparison my homelab security is very good. I am not a security expert nor did I try to dig deeper into security risks, because I think even trying to exploit them for this project would be illegal, but from what I've seen the way technology has advanced and the "make it work" mentality from what I gathered in my work and web related circles, it is much more common to have security as an afterthought until you get a serious attack on your network/infrastructure.

In regards to the python/streamlit programming part I learned a bit more how to better structure the code and the results. At the beginning I was not entirely sure what direction this project should take but the discussion and feedback helped me visualise it a bit better.

At the end of the project I forgot why I was using UV exactly, I had to reinstall my computer and just redid it with VENV and requirements.txt I do feel I should try to learn UV better in the future. Also I should try to integrate a linter/formatter as well I think (currently working with whatever PyCharm suggests).