FINAL REPORT

For

CLASS PROJECT

ANALYSING BENIGN & MALICIOUS WEBPAGES

Version 1.0

GROUP 03

GROUP MEMBERS

Kesavi Aravinthan - 190049P

Kajanan Selvanesan - 190287R

❖ Overview of the Project

There are so many web pages in this digital era. However, some of them are not benign. Visiting such malicious web pages can cause privacy and security issues. Therefore it is better to figure them out before visiting. Having an awareness about such web pages will save time and help to improve web security and protect users from threats.

Therefore, in this project, we have derived valuable insights from the selected dataset and presented those insights in an interactive dashboard.

Dataset: Dataset of Malicious and Benign Webpages [1]

See [2] and [3] for similar works that have been done in the same domain. However, They had mainly focused on model building. But here, insights and visualisations have been focused more.

❖ Objectives of the Project

We developed an interactive dashboard to analyse malicious and benign web pages. Therefore, the main deliverable of this project is

- ➔ Provide valuable insights and common patterns of malicious and benign web pages with interactive visualisation.

❖ The Used Tools

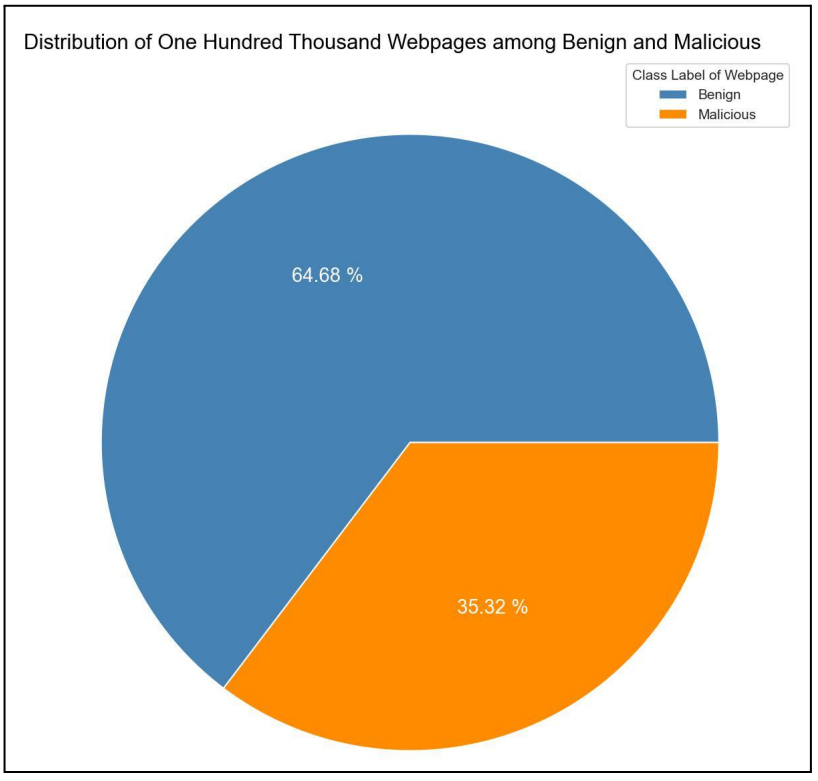
Tools	Where
Python	For data preprocessing & analysis, and visualisations
Tableau	To build an interactive dashboard

❖ Insights

We have mainly answered four (4) analytical questions using some insights. Those questions are:

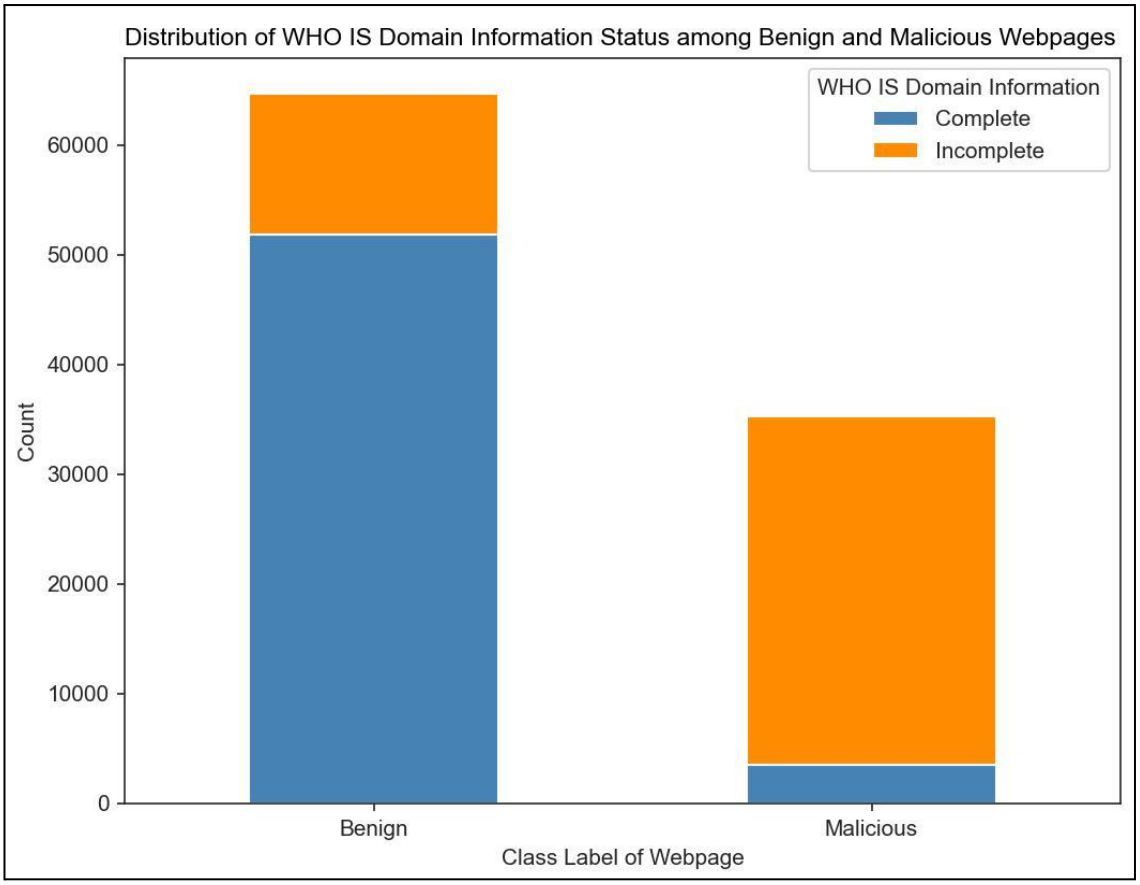
1. Can we decide whether a web page is malicious or not based on
    a. **WHO IS** domain information (Whether the WHO IS domain information is complete or not)
    b. **Internet Protocol** (Whether the site uses HTTPS or HTTP)
2. What are the typical **Top Level Domains** used for malicious web pages?
3. What are the typical **Geographic Locations** where the malicious web pages are hosted?
4. Can we decide whether a web page is malicious or not based on the **Length of the Content** on the web page?

We have done descriptive and exploratory data analysis on 100000 web pages of information. Out of those 100000 web pages, 64685 web pages are benign and 35315 web pages are malicious. In other words, We have used around 65% benign web pages and 35% malicious web pages in our analysis.



1. Can we decide whether a web page is malicious or not based on

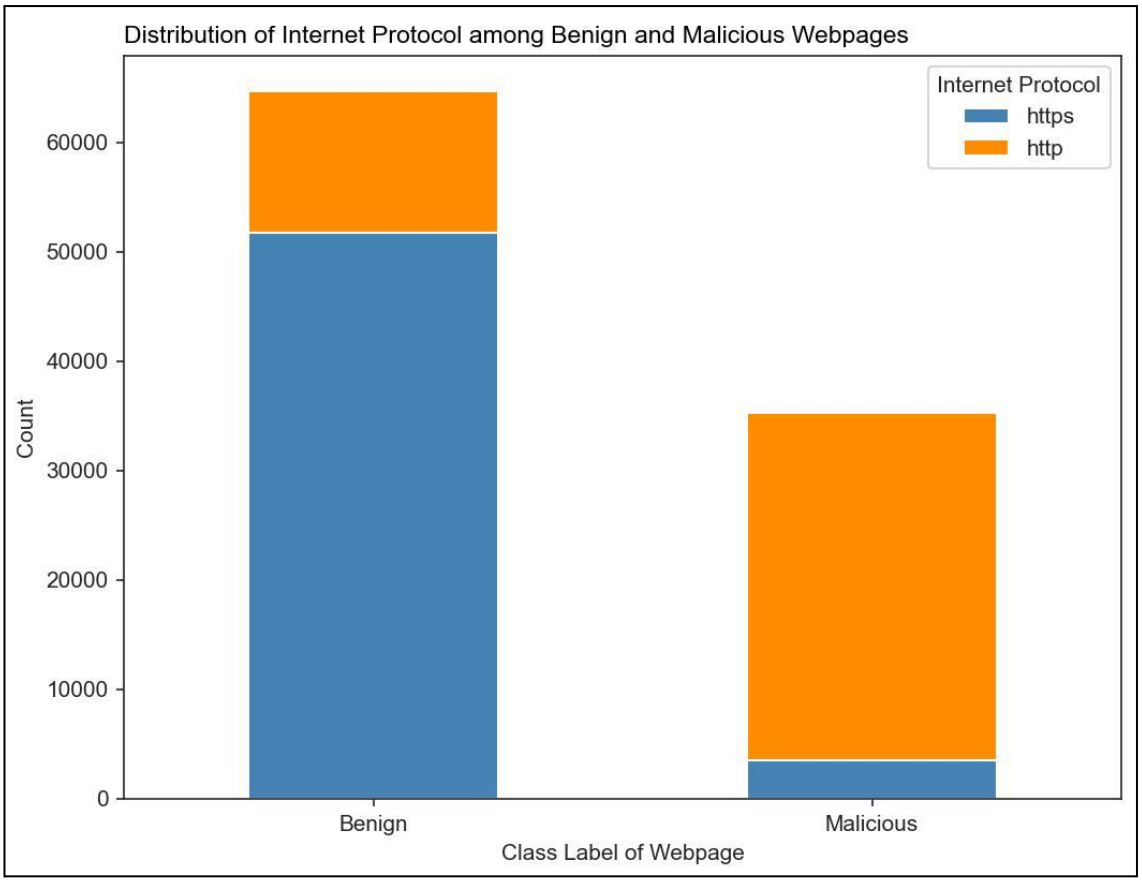
    a. WHO IS



The above diagram clearly shows that we can decide whether a web page is malicious or not based on whether the WHO IS domain information is complete or not. As per the diagram, a large portion of the benign web pages has complete WHO IS domain information, whereas only a tiny portion of the malicious web pages has complete WHO IS domain information.

As a result, we can say that if a web page has complete WHO IS domain information, then that web page has a high probability of being a benign webpage. In contrast, if a web page has incomplete WHO IS domain information, then that web page has a high probability of being a malicious webpage.

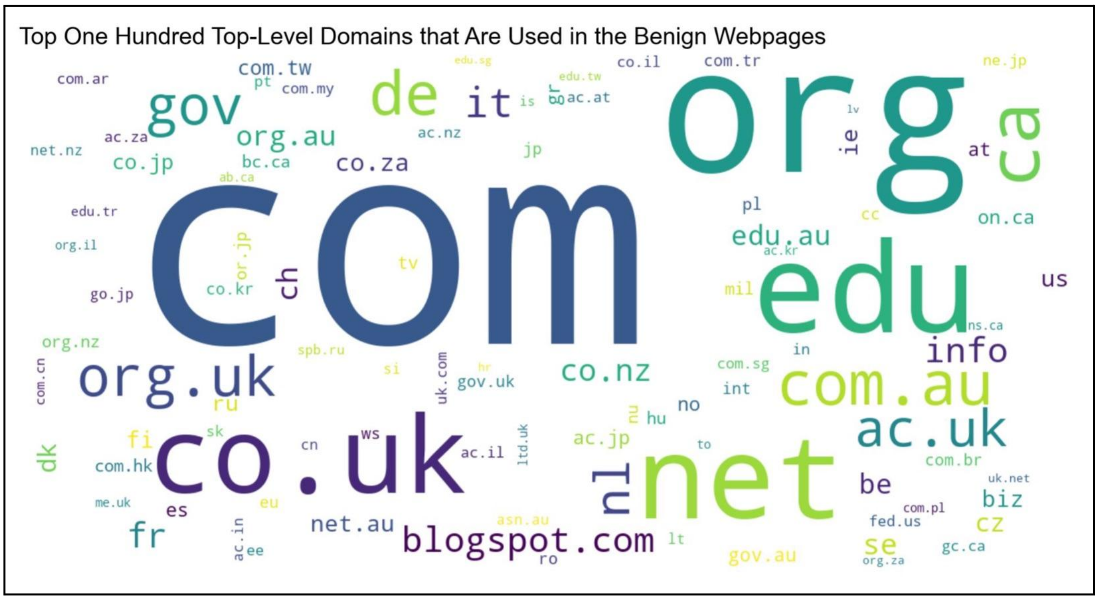
b. Internet Protocol

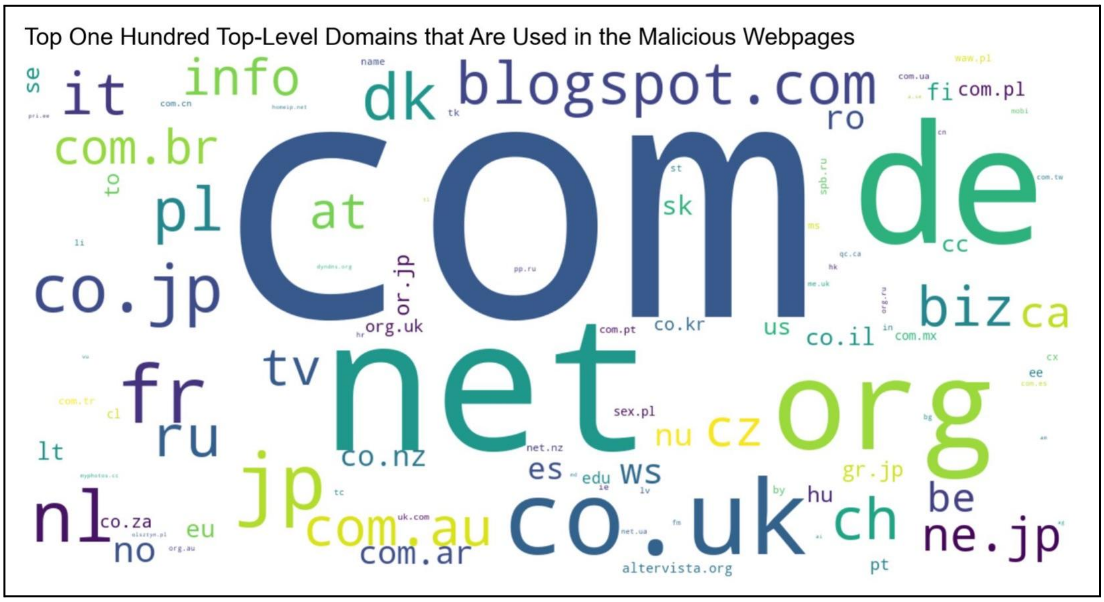


The above diagram clearly shows that we can decide whether a web page is malicious or not based on whether the site uses HTTPS or HTTP. As per the diagram, a large portion of the benign web pages has Hypertext Transfer Protocol Secure, whereas only a tiny portion of the malicious web pages has Hypertext Transfer Protocol Secure.

As a result, we can say that if a web page has Hypertext Transfer Protocol Secure, then that web page has a high probability of being a benign webpage. In contrast, if a web page has Hypertext Transfer Protocol Secure, then that web page has a high probability of being a malicious webpage.

2. What are the typical Top Level Domains used for malicious web pages?



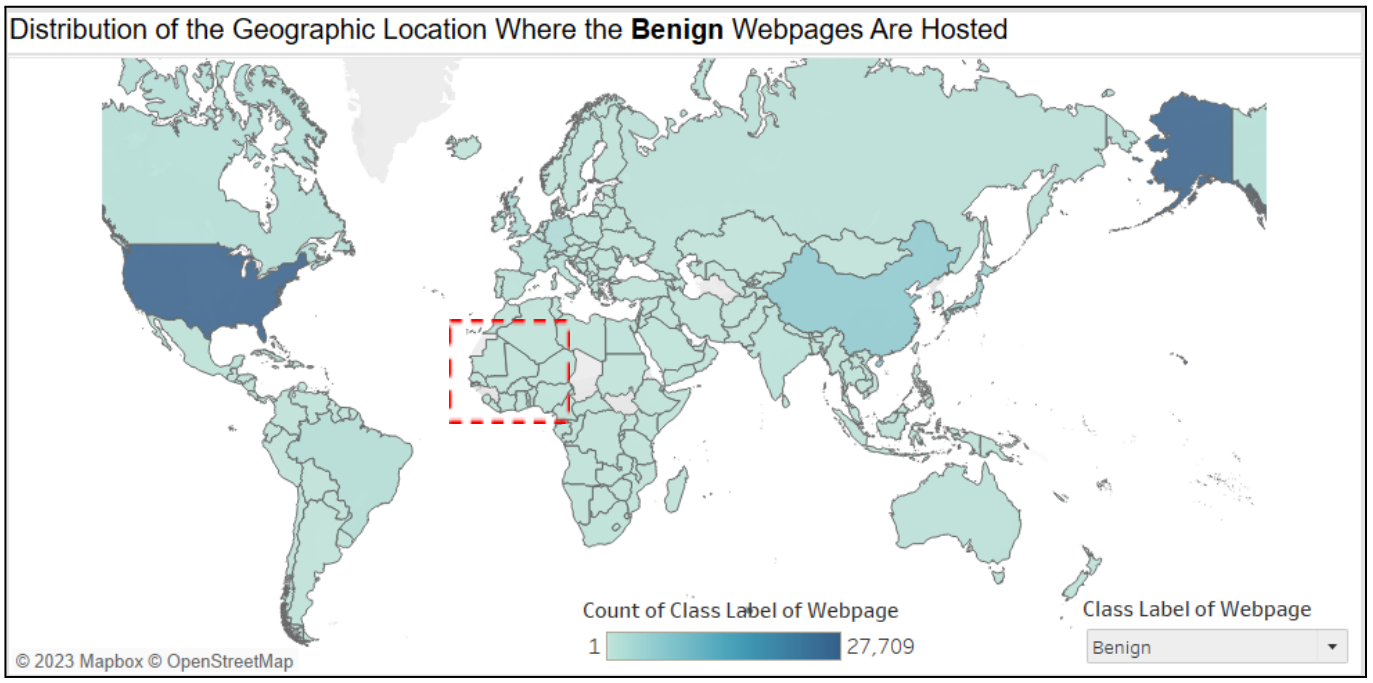


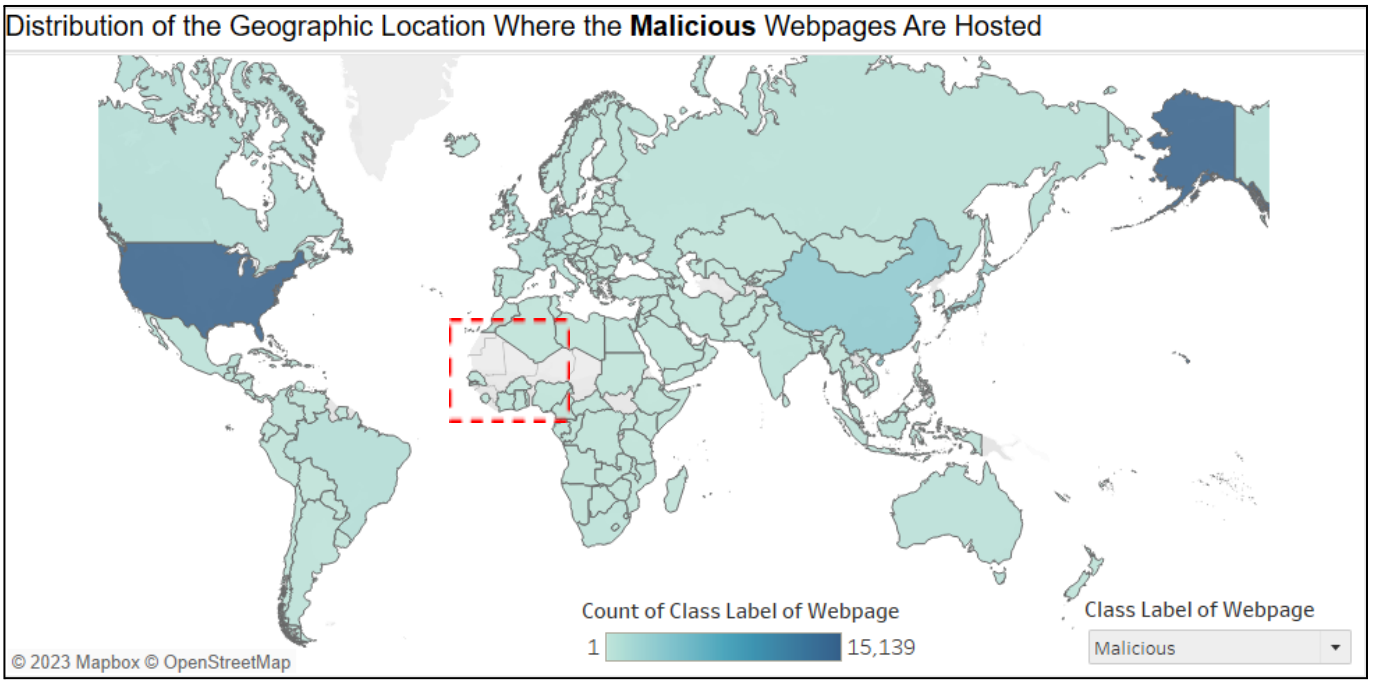
The above diagram clearly shows that some Top Level Domains are highly used for malicious web pages. For example: **“de”**, **“dk”**, **“ru”**, **“tv”** and **“jp”**.

And some Top Level Domains are highly used for benign web pages. For example: **“edu”** and **“gov”**. These are obvious since **“edu”** is used for organisations which focus on education, and **“gov”** is used by government entities.

However, some Top Level Domains like **“com”**, **“org”** and **“net”**, are widely used in both types of web pages.

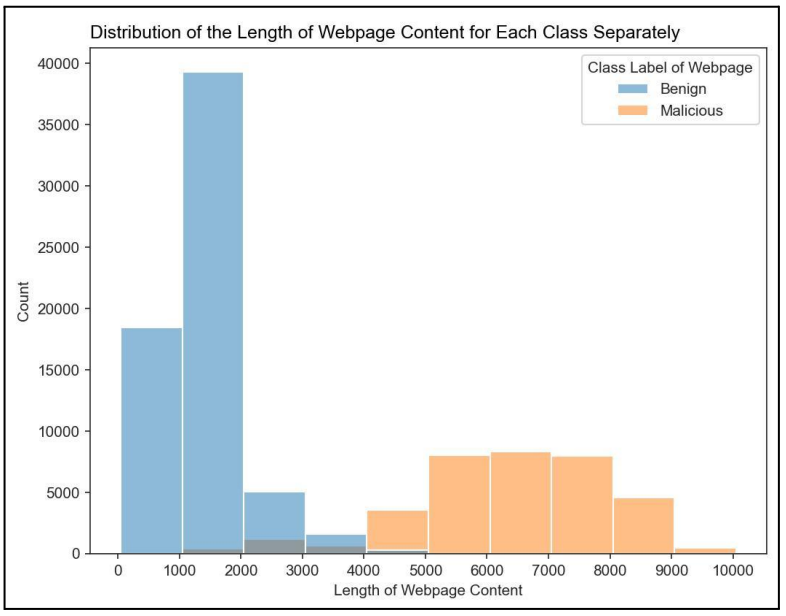
3. What are the typical Geographic Locations where the malicious web pages are hosted?





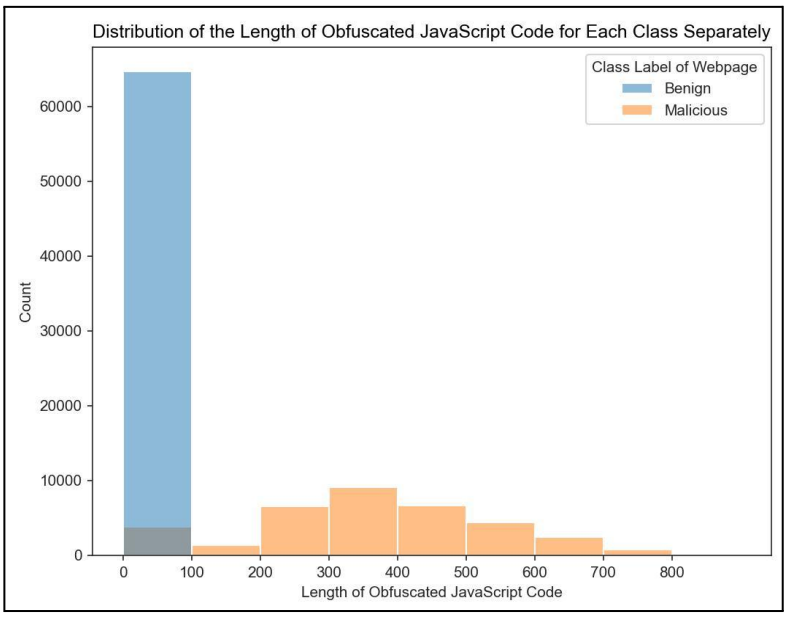
From the above diagrams, we can not come to a conclusion that these are the typical Geographic Locations where malicious web pages are hosted. But we can say that the square area that is outlined as red, is not typically used to host the malicious web pages.

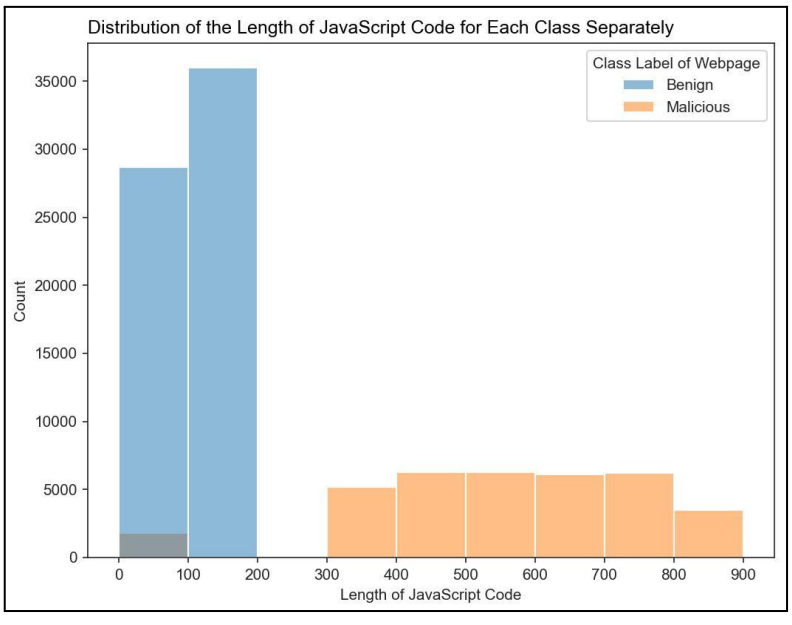
4. Can we decide whether a web page is malicious or not based on the length of the content in the web page?



The above diagram clearly shows that we can decide whether a web page is malicious or not based on the length of the content on the web page. As per the diagram, the typical range for the length of the content in benign web pages is less than that of malicious web pages.

As a result, we can say that if a web page has more than 5000 words in its content, then that web page has a very high probability of being a malicious webpage. In contrast, if a web page has less than 4000 words in its content, then that web page has a high probability of being a benign webpage.





In the above two diagrams also, we can notice similar behaviours like the behaviour of length of the content on the web page. We can decide whether a web page is malicious or not based on the length of the JavaScript or obfuscated JavaScript code that is used in the web page. As per the diagram, the typical range for the length of the JavaScript or obfuscated JavaScript code that is used in benign web pages is less than that of malicious web pages.

As a result, we can say that if a web page has more than 300 lengths of JavaScript code, then that web page has a very high probability of being a malicious webpage. In contrast, if a web page has less than 200 lengths of JavaScript code, then that web page has a high probability of being a benign webpage.

Similarly, we can say that if a web page has more than 100 lengths of obfuscated JavaScript code, then that web page has a very high probability of being a malicious webpage. In contrast, if a web page has less than 100 lengths of obfuscated JavaScript code, then that web page has a high probability of being a benign webpage.

❖ References

[1] Singh, AK (2020), "Dataset of Malicious and Benign Webpages", Mendeley Data, V2, doi: 10.17632/gdx3pkwp47.2

[2] Mamun, M.S.I., Rathore, M.A., Habibi Lashkari, A., Stakhanova, N., & Ghorbani, A.A. (2016). Detecting Malicious URLs Using Lexical Analysis. In: Chen, J., Piuri, V., Su, C., Yung, M. (eds) Network and System Security. NSS 2016. Lecture Notes in Computer Science(), vol 9955. Springer, Cham. https://doi.org/10.1007/978-3-319-46298-1_30

[3] Alaa El-khashap. 2022. Malicious-URLs. https://github.com/alaaelkhashap/Malicious-URLs. (2023).