Nithin S
221IT085

IT352 Course Project Module 1

Verifying existing URLs with VirusTotal and Extracting Features from URL Dataset to build a new dataset

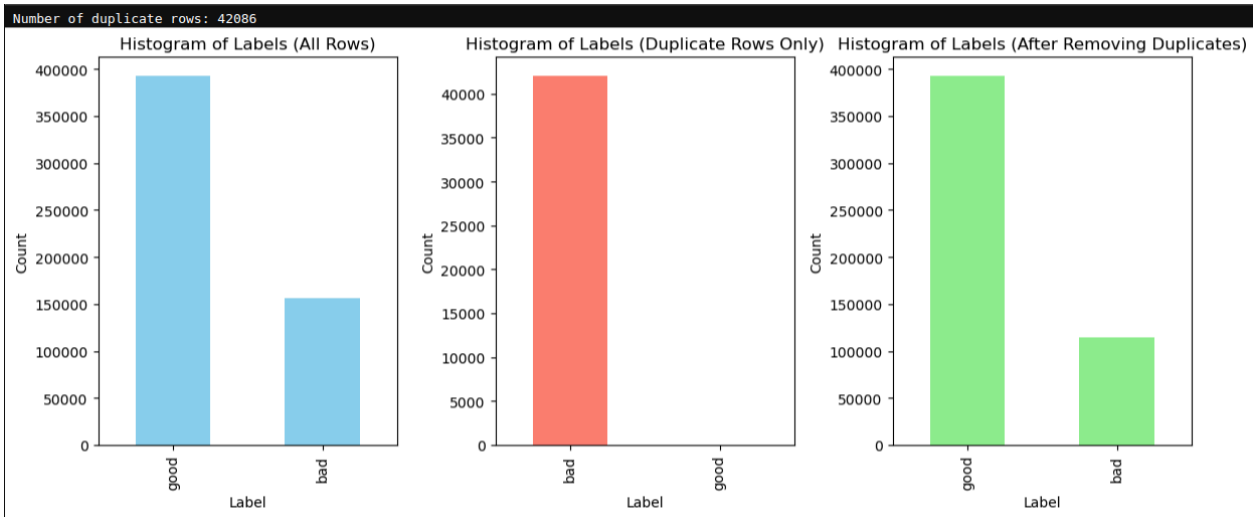
Phishing Site URLs: Dataset which contains Phishing urls and non phishing urls.

Total URLs: 549362
Total Unique URLs: 507195

URL unique URLs	Label good and bad URLs classes
507195 **unique values**	**good** 72% **bad** 28%
nobell.it/70ffb52d079109dca5664cce6f317373782/login.SkyPe.com/en/cgi-bin/verification/login/70ffb52d...	bad
www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrcmd=_home-customer&nav=1/loading.php	bad
serviciosbys.com/paypal.cgi.bin.get-into.herf.secure.dispatch35463256rzr321654641dsf654321874/href/h...	bad

Histogram about the classes in the URLs Dataset

Total Rows present in given dataset: 549346
No of duplicates : 42086
Percentage of duplicates: 7.66 %
Total Unique URLs: 507195



No Null Values were found in the given URLs Dataset

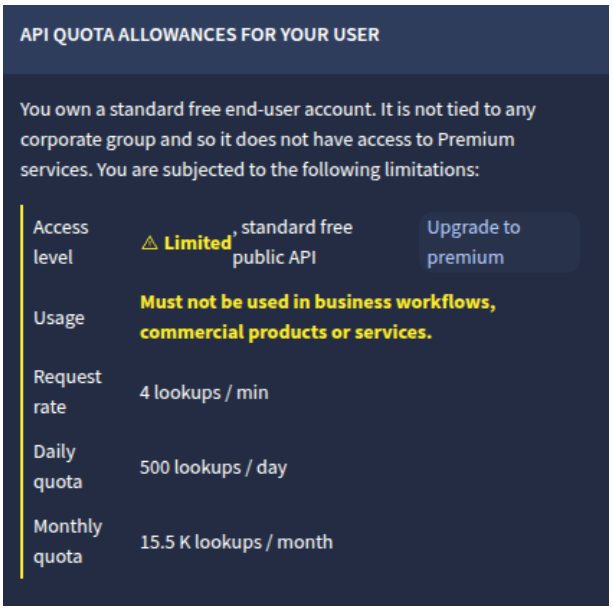
```
phish_data.isnull().sum() # there is no missing values

URL      0
Label    0
dtype: int64
```

Python Script to verify the labelling of the dataset

Please make sure to add required API Keys before running the code.

Virus Total API : https://docs.virustotal.com/reference/scan-url

Virus Total API Endpoint has strict rate limits



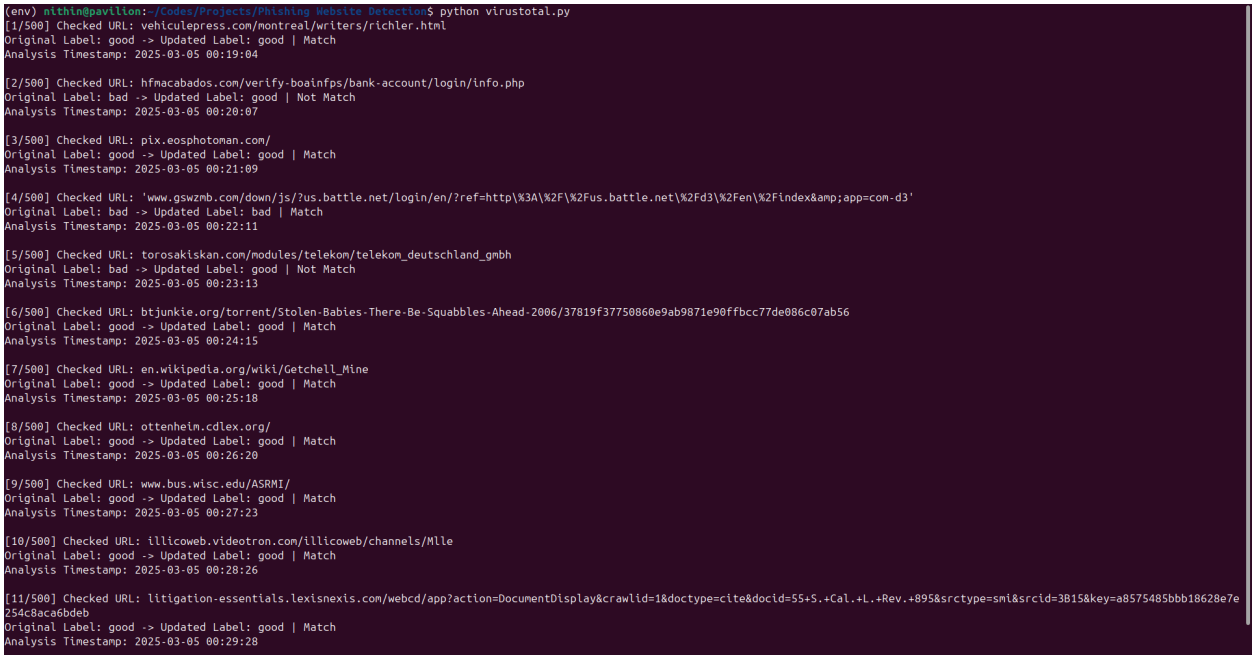
So, we will verify random 500 urls from the dataset without repetition and update them. We will have a **time delay of 60s** between each URL Verification request.

CODE

```
def check_url_virustotal(url, default_label):
    endpoint = "https://www.virustotal.com/api/v3/urls"
    headers = {"x-apikey": os.getenv("VIRUS_TOTAL_API_KEY")}
    try:
        response = requests.post(endpoint, headers=headers, data={"url": url})
        if response.status_code != 200:
            return default_label
        analysis_id = response.json()["data"]["id"]
        result_endpoint = f"https://www.virustotal.com/api/v3/analyses/{analysis_id}"
        result_response = requests.get(result_endpoint, headers=headers)
        if result_response.status_code != 200:
            return default_label
```

```
        result_data =
result_response.json()["data"]["attributes"]["results"]
        malicious_count = sum(1 for scan in result_data.values() if
scan["category"] == "malicious")
        return "bad" if malicious_count > 0 else "good"
    except Exception:
        return default_label
```

Output



Total mismatches : 120 out of 500
Percentage of mismatches: 120/500 * 100 = 24%

Due to size limitations I have uploaded the datasets in my google drive.

Extracted Features Dataset Google Drive Link:
https://drive.google.com/file/d/119L_eJDb8Oizm4jv3NM0tfW-Um-7sSDC/view?usp=sharing

Z Score Normalized Dataset Google Drive Link:
https://drive.google.com/file/d/1P8sOZi6j7JeM_LNLFmChU9FPHnyo_u1X/view?usp=sharing

List of Features planned to be extracted: 116 Features - (Without dropping repeated columns)

Full URL Length: Total number of characters in the entire URL.
Hostname (Domain) Length: Total number of characters in the domain name part.
Directory Length: Number of characters in the folder or path part of the URL.
File Name Length: Number of characters in the file name portion (if any).
Parameters Length: Number of characters in the query string (everything after “?”).
TLD Length: Length (in characters) of the top-level domain (for example, “com” or “org”).

Dot ('.') Count: Number of periods used (often separating subdomains or domain parts).
Hyphen ('-') Count: Number of hyphens used.
Underscore ('_') Count: Number of underscores.
Slash ('/') Count: Number of forward slashes.
Question Mark ('?') Count: How many “?” appear.
Equal Sign ('=') Count: Number of equal signs.
At Sign ('@') Count: Number of “@” symbols.
Ampersand ('&') Count: How many “&” symbols appear.
Exclamation Mark ('!') Count: Number of “!” symbols.
Space Count: Number of space characters.
Tilde ('~') Count: Number of tilde characters.
Comma (',') Count: How many commas appear.
Plus Sign ('+') Count: Number of “+” symbols.
Asterisk ('*') Count: Number of “*” symbols.
Hashtag ('#') Count: Number of “#” symbols.
Dollar Sign ('$') Count: Number of “$” symbols.
Percent Sign ('%') Count: Number of “%” symbols.

Common Terms Occurrence: Counts for terms such as “www”, “.com”, “http”, and “//” that usually appear only once in normal URLs.
Email in URL: A flag indicating if an email address is embedded in the URL.

HTTPS Token: Checks if the URL uses “https” (a sign of secure connections).
IP Address in URL: A binary check to see if an IP address is used instead of a domain name.
Punycode Usage: Checks whether the domain uses punycode (which can mask its true characters).
Port Number Presence: A flag indicating if the URL explicitly shows a port (like “:80” or “:443”).
TLD Position: Verifies that the top-level domain is in the right place (it should not appear in the wrong section like the path or subdomain).

Abnormal Subdomains: Detects unusual subdomain patterns (for example, variations of "www" that include numbers).
Number of Subdomains: Counts how many subdomains are present.
Prefix/Suffix with Hyphen: Checks if the domain uses hyphens to separate extra words (which might be used to mimic legitimate sites).
Random Domain Indicator: Determines if the domain seems to be made up of random characters.
URL Shortening Service: A flag to see if a URL shortener (like bit.ly) is used, which can hide the true destination.
Path Extension Check: Looks for suspicious file extensions (such as ".exe" or ".js") in the URL path.
Suspicious TLD: Checks if the top-level domain is among those known to be risky.

Digit Ratio in Full URL: Proportion of digit characters compared to the total characters in the URL.
Digit Ratio in Hostname: Proportion of digits in the domain name itself.

Word Count: Number of words found in the full URL, the hostname, or the path.
Shortest & Longest Word: Identification of the shortest and longest word in the URL parts.
Average Word Length: The average length of words in the URL, hostname, or path.
Phish Hints: Counts occurrences of suspicious or phishing-related keywords (like "login", "admin", "signin", etc.).

Brand Names in URL:
In the Domain: Presence of well-known brand names can be a sign of legitimacy.
In the Subdomain or Path: Their appearance here may indicate an attempt to deceive.
Domain in Page Title/Copyright: Checks if the domain name appears in the webpage title or copyright text (a sign of legitimacy).

Redirection Count: Total number of times the URL redirects to another page.
External Redirections: How many of these redirects go to a different domain.
Internal vs. External Hyperlinks Ratio: Compares links that point within the same site to those that point to external sites.
Null Hyperlinks Ratio: Proportion of links that lead nowhere (empty links).
Media Links Ratio: Ratio of media (images, videos, etc.) hosted on the same domain versus externally.
Connection Errors Ratio: Ratio of hyperlinks that result in errors (broken links).

Number of Hyperlinks: Total links present on the webpage.
External CSS Files Count: Number of CSS files linked from outside the domain.
Login Forms Presence: Checks for login forms, especially those with empty or suspicious action attributes.
External Favicon: Whether the page uses a favicon (the small icon in the browser tab) from an external source.

Invisible iFrame: Detects hidden iframe elements that might load content from another domain.
Pop-up Windows: Looks for pop-up windows that include text fields (which can be a sign of phishing).
Unsafe Anchors: Counts anchor (<a>) tags that use unsafe links (e.g., “javascript:” or “#”).
Right-Click Blocking: Checks for scripts that disable the right-click function (which can hide page source).
Empty Title: Flags if the webpage has no title tag.

WHOIS Registration: Whether the domain is found in the WHOIS database (a missing record is a red flag).
Domain Registration Length: The number of years for which the domain is registered (short registration periods can be suspicious).
Domain Age: How long the domain has been active.
DNS Record Check: Verifies that the domain has proper DNS records.
Google Index: Checks if the URL or domain is indexed by Google (phishing sites are often not).
Page Rank: An estimate of the webpage’s popularity.
Web Traffic: An indicator (like Alexa ranking) showing the number of visitors.

Additionally, one study mentions a “statistical report” feature that checks if the domain’s IP matches known top phishing domains.

Vowel Count in Domain: Number of vowels in the domain name.
Domain in IP Format: Whether the domain is written as an IP address.
“Server” or “Client” in Domain: Checks if these words appear in the domain name, which can hint at its purpose.

Domain Lookup Response Time: How long it takes to get a response when looking up the domain.
SPF Record: Checks if the domain has an SPF record (helps validate email sources).
ASN (Autonomous System Number): A number that identifies the network the domain’s IP belongs to.
Domain Activation Time: How many days have passed since the domain was first activated.
Domain Expiration Time: How many days remain until the domain expires.
Number of Resolved IPs: How many IP addresses are returned when the domain is looked up.
Nameservers Count: Number of DNS nameservers linked to the domain.
MX Servers Count: Number of mail servers associated with the domain.
TTL (Time-To-Live) of Hostname: The DNS record’s lifetime.
Valid TLS/SSL Certificate: Whether the site has a proper secure certificate.
URL Shortened Flag: Whether the URL has been shortened (also noted earlier under security).

TLD Present in Parameters: Checks if a top-level domain appears within the URL parameters (which is unusual).
Number of Parameters: Count of key–value pairs or parameters present in the URL query string.

Check for duplicate Columns

```
Column Index: 14, Column Name: tilde_count
Column Index: 28, Column Name: https_token
Column Index: 60, Column Name: brand_in_subdomain
Column Index: 86, Column Name: whois_registration
Column Index: 87, Column Name: domain_registration_length
Column Index: 88, Column Name: domain_age
Column Index: 96, Column Name: server_or_client_in_domain
Column Index: 98, Column Name: asn
Column Index: 99, Column Name: domain_activation_time
Column Index: 100, Column Name: domain_expiration_time
```

Check for duplicate rows

```
5339
9567
16871
18507
18965
18967
19005
19398
19673
19897
22329
22413
22605
22755
22762
22824
22892
```

Normalization

I have normalized all columns , since all were numerical.

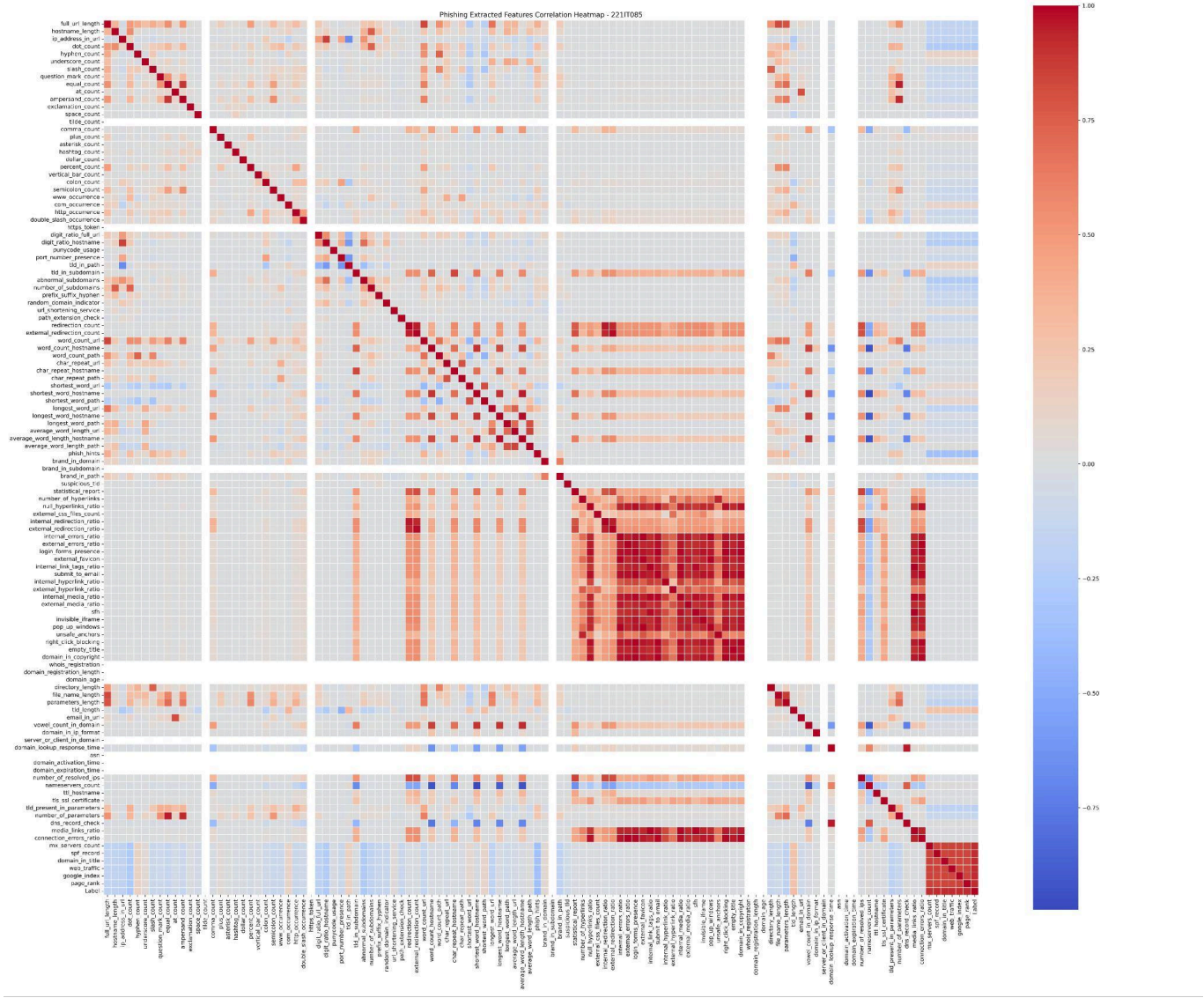
```
Standardization (Z-score normalization) applied to: ['full_url_length', 'hostname_length', 'ip_address_in_url', 'dot_count', 'hyphen_count', 'underscore_count', 'slash_count', 'question_mark_count', 'equal_count', 'at_c
ount', 'ampersand_count', 'exclamation_count', 'space_count', 'tilde_count', 'comma_count', 'plus_count', 'asterisk_count', 'hashtag_count', 'dollar_count', 'percent_count', 'vertical_bar_count', 'colon_count', 'semicol
on_count', 'www_occurrence', 'com_occurrence', 'http_occurrence', 'double_slash_occurrence', 'https_token', 'digit_ratio_full_url', 'digit_ratio_hostname', 'punycode_usage', 'port_number_presence', 'tld_in_path', 'tld_i
n_subdomain', 'abnormal_subdomains', 'number_of_subdomains', 'prefix_suffix_hyphen', 'random_domain_indicator', 'url_shortening_service', 'path_extension_check', 'redirection_count', 'external_redirection_count', 'word_
count_url', 'word_count_hostname', 'word_count_path', 'char_repeat_url', 'char_repeat_hostname', 'char_repeat_path', 'shortest_word_url', 'shortest_word_hostname', 'shortest_word_path', 'longest_word_url', 'longest_word
_hostname', 'longest_word_path', 'average_word_length_url', 'average_word_length_hostname', 'average_word_length_path', 'phish_hints', 'brand_in_domain', 'brand_in_subdomain', 'brand_in_path', 'suspicious_tld', 'statist
ical_report', 'number_of_hyperlinks', 'null_hyperlinks_ratio', 'external_css_files_count', 'internal_redirection_ratio', 'external_redirection_ratio', 'internal_errors_ratio', 'external_errors_ratio', 'login_forms_prese
nce', 'external_favicon', 'internal_link_tags_ratio', 'submit_to_email', 'internal_hyperlink_ratio', 'external_hyperlink_ratio', 'internal_media_ratio', 'external_media_ratio', 'sfh', 'invisible_iframe', 'pop_up_window
s', 'unsafe_anchors', 'right_click_blocking', 'empty_title', 'domain_in_copyright', 'whois_registration', 'domain_registration_length', 'domain_age', 'directory_length', 'file_name_length', 'parameters_length', 'tld_len
gth', 'email_in_url', 'vowel_count_in_domain', 'domain_in_ip_format', 'server_or_client_in_domain', 'domain_lookup_response_time', 'asn', 'domain_activation_time', 'domain_expiration_time', 'number_of_resolved_ips', 'na
meservers_count', 'ttl_hostname', 'tls_ssl_certificate', 'tld_present_in_parameters', 'number_of_parameters', 'dns_record_check', 'media_links_ratio', 'connection_errors_ratio', 'mx_servers_count', 'spf_record', 'domain
_in_title', 'web_traffic', 'google_index', 'page_rank']
```

Handle Missing Values

```
df.isnull().sum()

URL                  0
full_url_length      0
hostname_length      0
ip_address_in_url    0
dot_count            0
                    ..
domain_in_title      0
web_traffic          0
google_index         0
page_rank            0
Label                0
Length: 117, dtype: int64
```

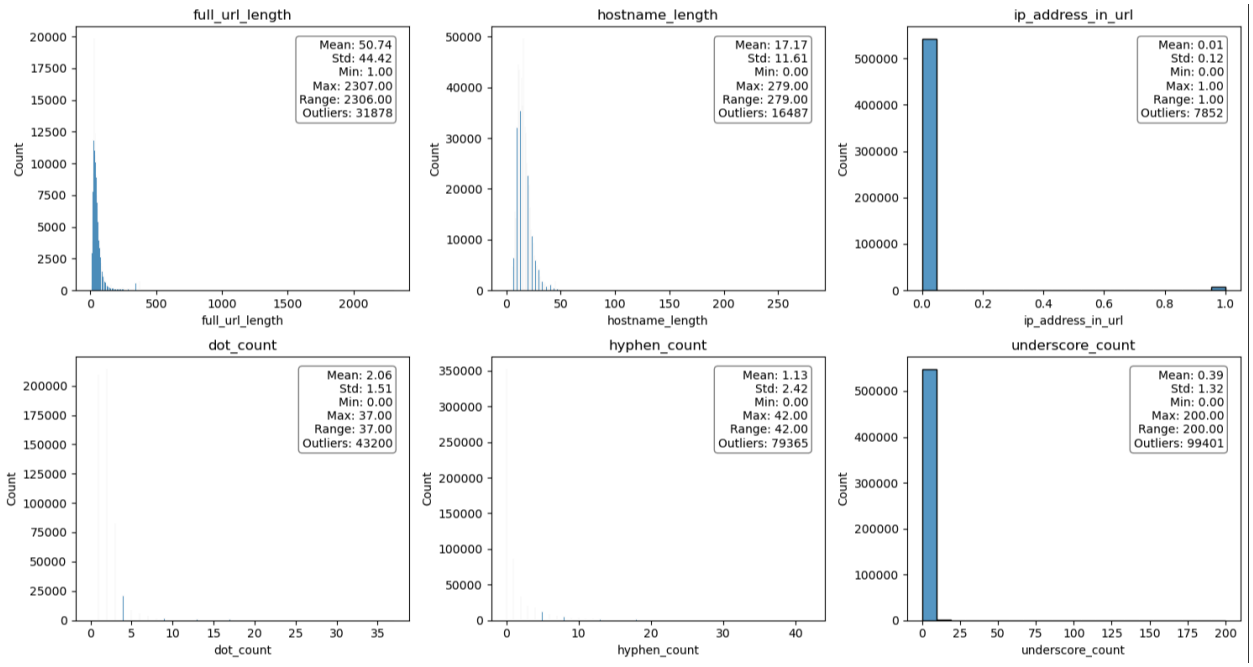
NO missing values were found

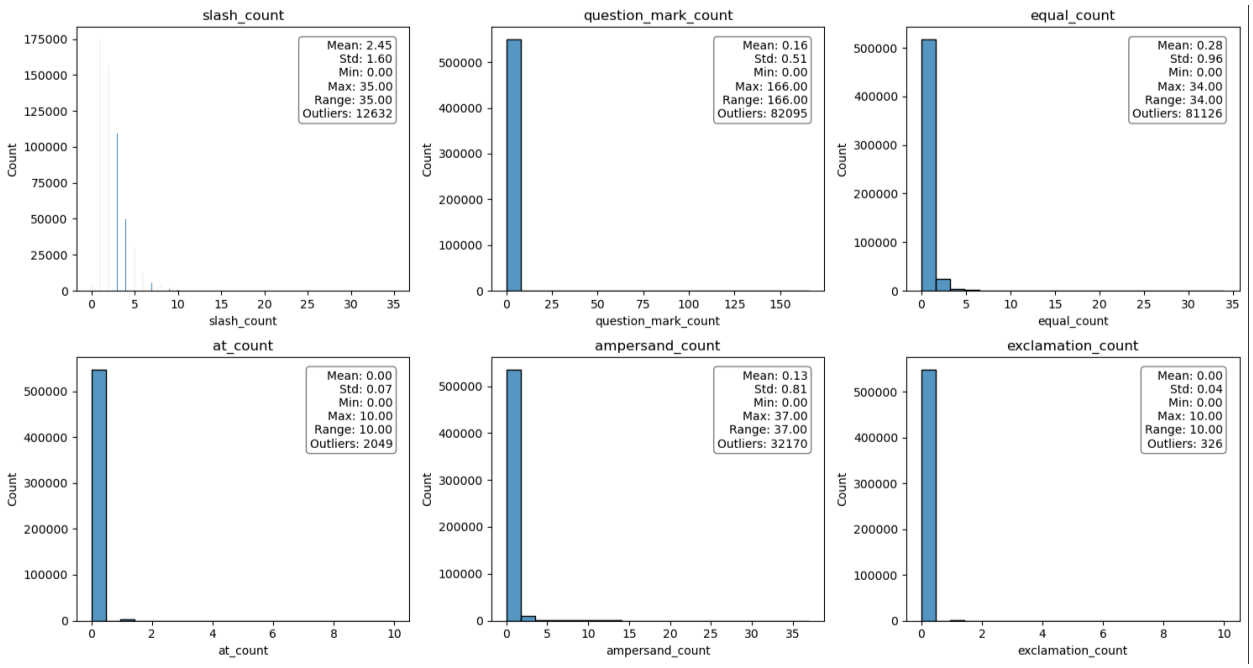
Correlation Map

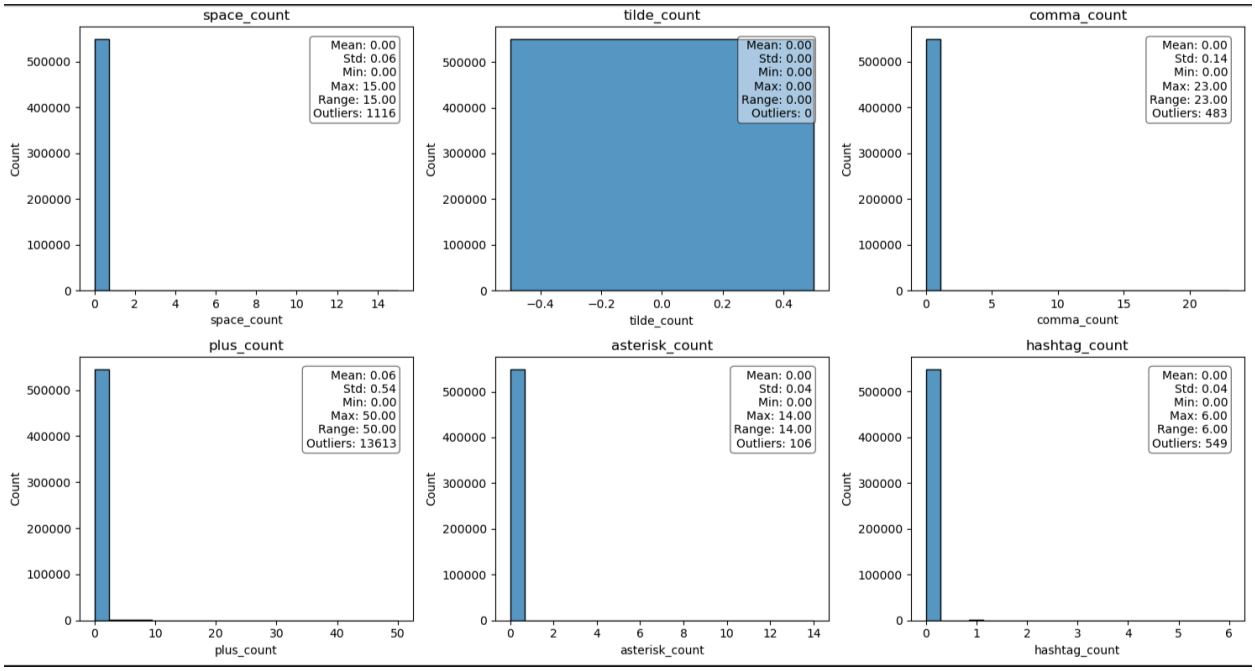


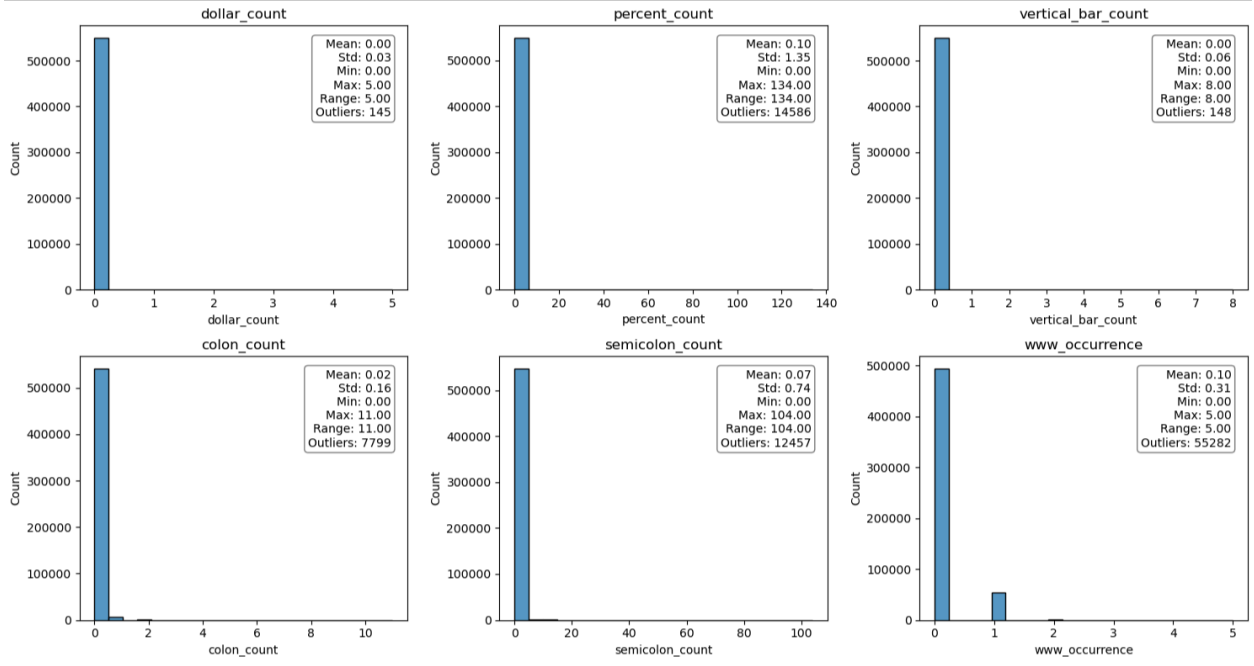
Please download the jpeg image from moodle and zoom into it for better view.

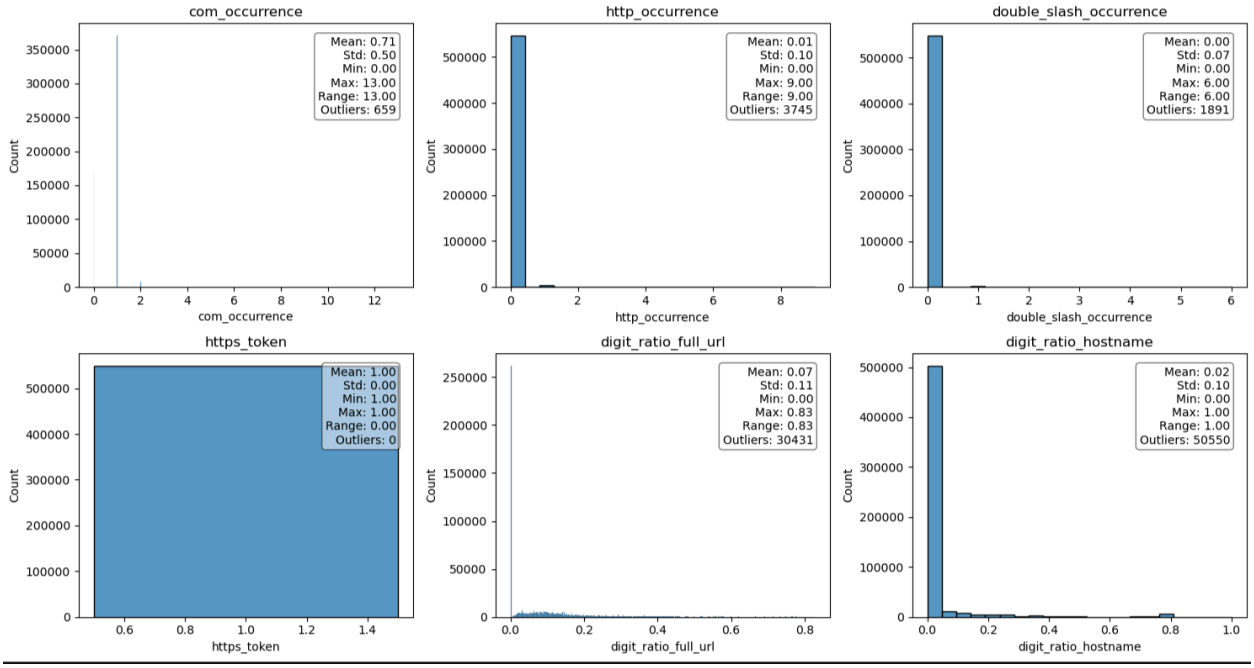
Distribution of Each Feature with its mean, standard deviation, min, max, range and outlier count

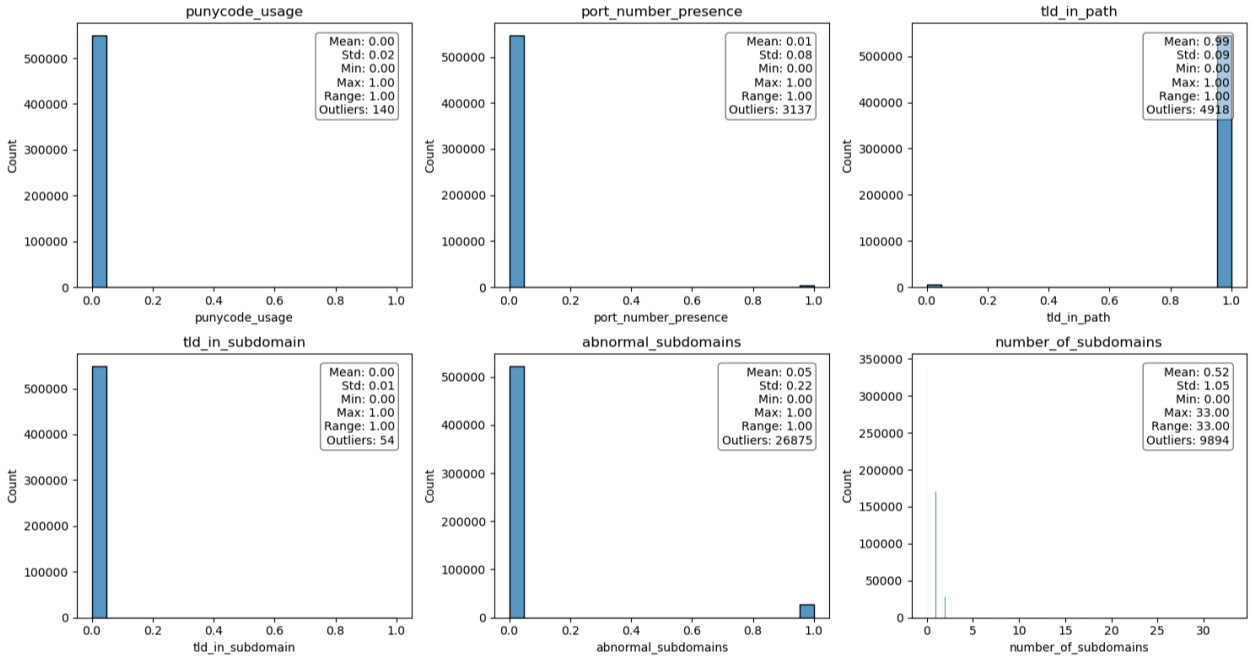


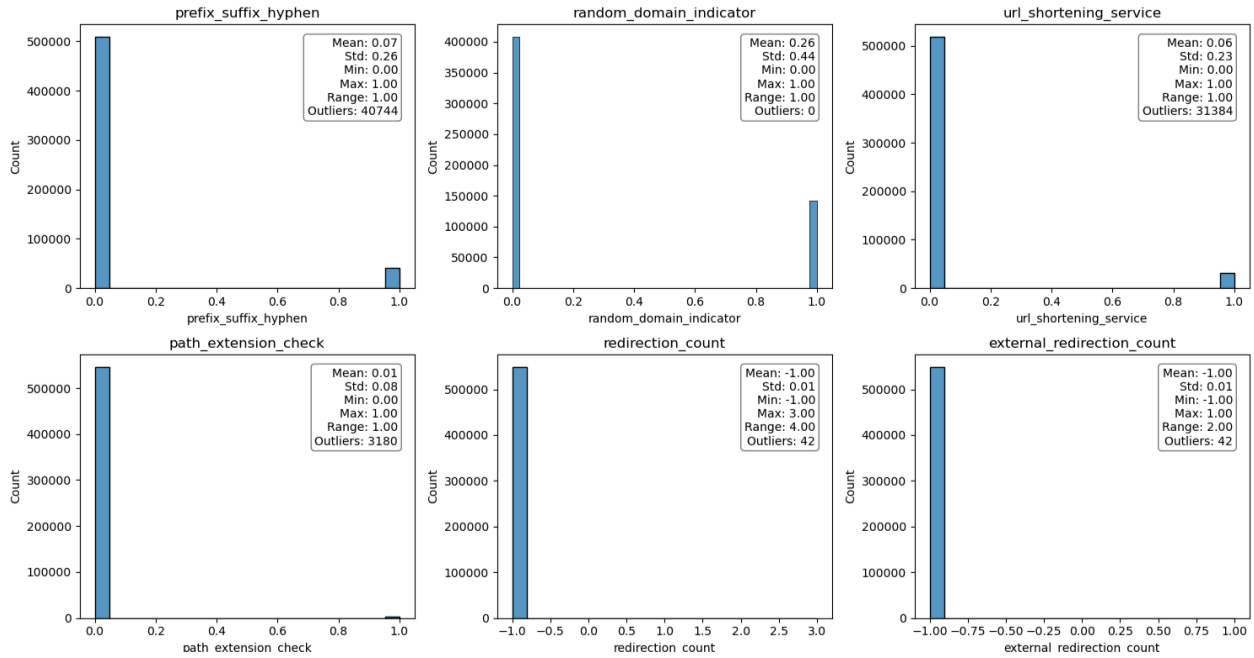


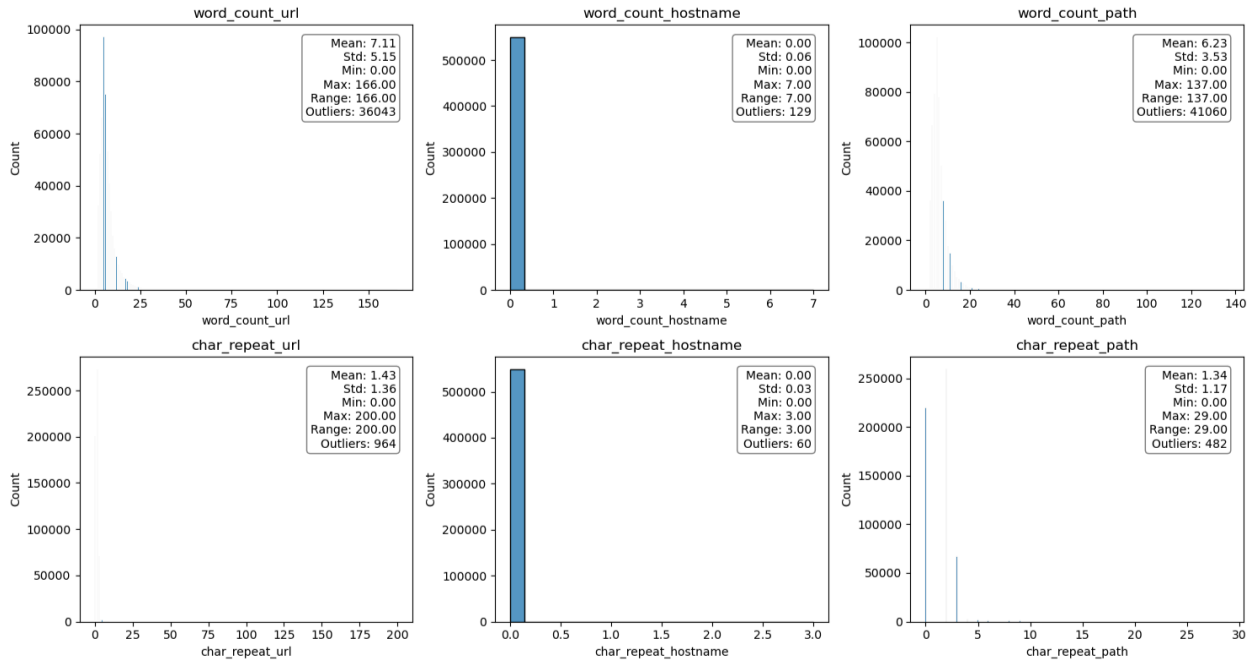

space_count
Mean: 0.00
Std: 0.06
Min: 0.00
Max: 15.00
Range: 15.00
Outliers: 1116
tilde_count
Mean: 0.00
Std: 0.00
Min: 0.00
Max: 0.00
Range: 0.00
Outliers: 0
comma_count
Mean: 0.00
Std: 0.14
Min: 0.00
Max: 23.00
Range: 23.00
Outliers: 483
plus_count
Mean: 0.06
Std: 0.54
Min: 0.00
Max: 50.00
Range: 50.00
Outliers: 13613
asterisk_count
Mean: 0.00
Std: 0.04
Min: 0.00
Max: 14.00
Range: 14.00
Outliers: 106
hashtag_count
Mean: 0.00
Std: 0.04
Min: 0.00
Max: 6.00
Range: 6.00
Outliers: 549
Count

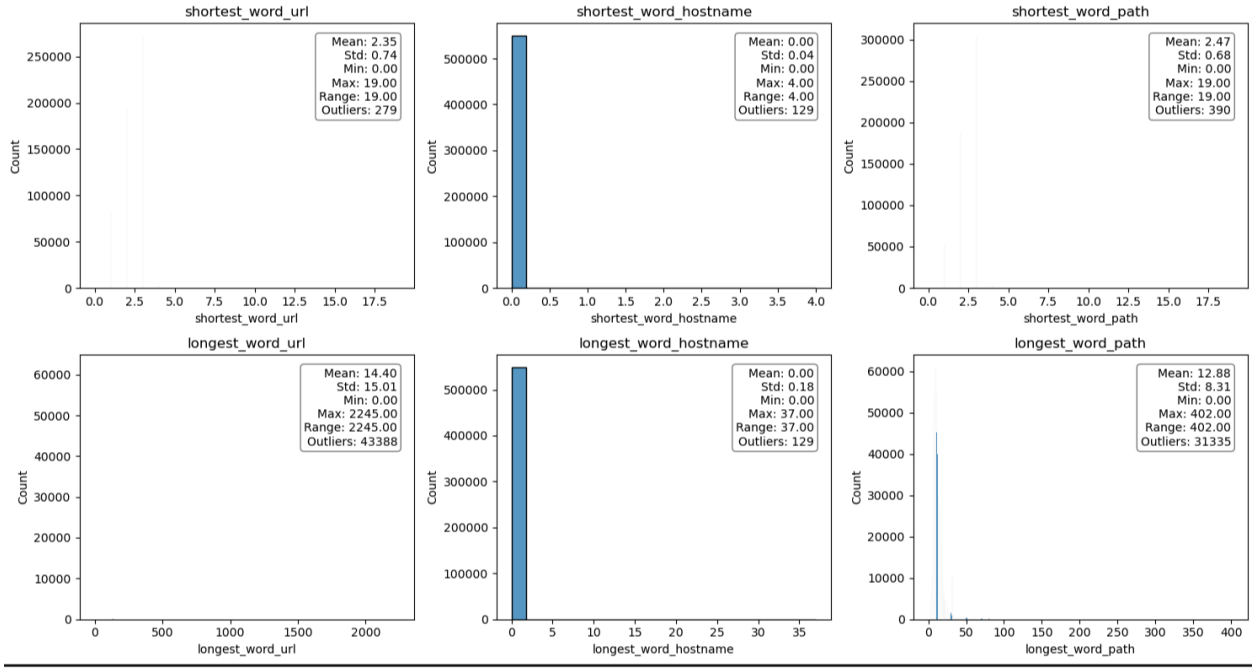

dollar_count
Mean: 0.00
Std: 0.03
Min: 0.00
Max: 5.00
Range: 5.00
Outliers: 145
percent_count
Mean: 0.10
Std: 1.35
Min: 0.00
Max: 134.00
Range: 134.00
Outliers: 14586
vertical_bar_count
Mean: 0.00
Std: 0.06
Min: 0.00
Max: 8.00
Range: 8.00
Outliers: 148
colon_count
Mean: 0.02
Std: 0.16
Min: 0.00
Max: 11.00
Range: 11.00
Outliers: 7799
semicolon_count
Mean: 0.07
Std: 0.74
Min: 0.00
Max: 104.00
Range: 104.00
Outliers: 12457
www_occurrence
Mean: 0.10
Std: 0.31
Min: 0.00
Max: 5.00
Range: 5.00
Outliers: 55282
Count

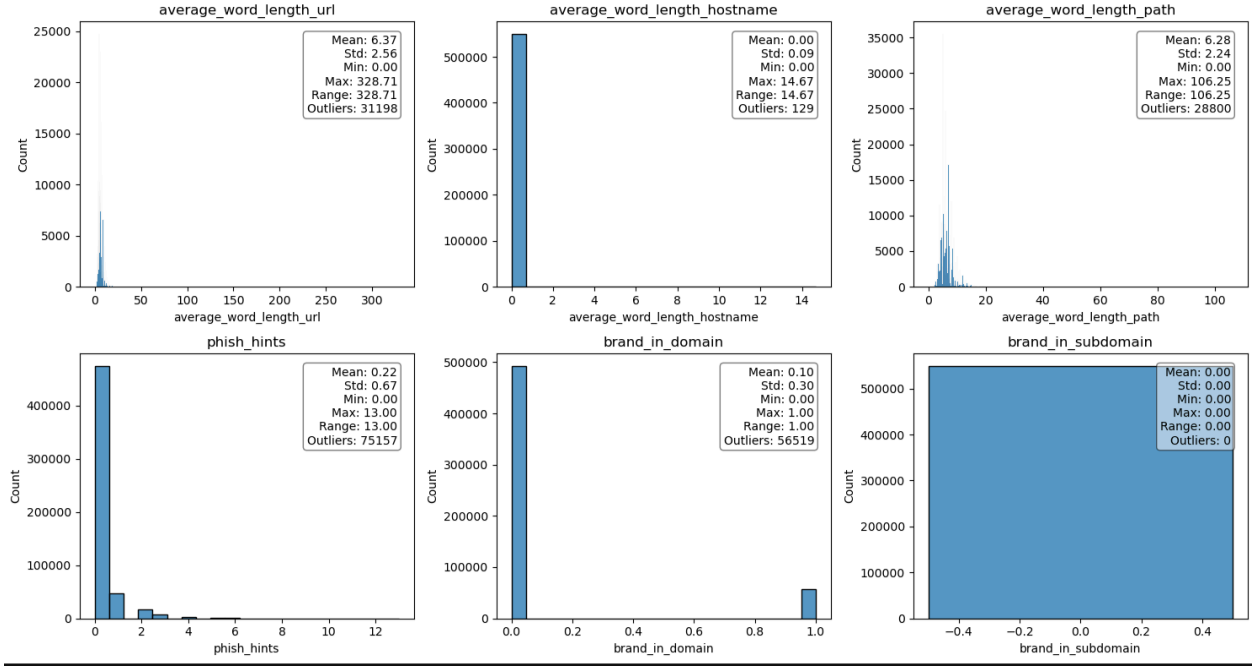

com_occurrence
Mean: 0.71
Std: 0.50
Min: 0.00
Max: 13.00
Range: 13.00
Outliers: 659
Count
com_occurrence
http_occurrence
Mean: 0.01
Std: 0.10
Min: 0.00
Max: 9.00
Range: 9.00
Outliers: 3745
Count
http_occurrence
double_slash_occurrence
Mean: 0.00
Std: 0.07
Min: 0.00
Max: 6.00
Range: 6.00
Outliers: 1891
Count
double_slash_occurrence
https_token
Mean: 1.00
Std: 0.00
Min: 1.00
Max: 1.00
Range: 0.00
Outliers: 0
Count
https_token
digit_ratio_full_url
Mean: 0.07
Std: 0.11
Min: 0.00
Max: 0.83
Range: 0.83
Outliers: 30431
Count
digit_ratio_full_url
digit_ratio_hostname
Mean: 0.02
Std: 0.10
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 50550
Count
digit_ratio_hostname

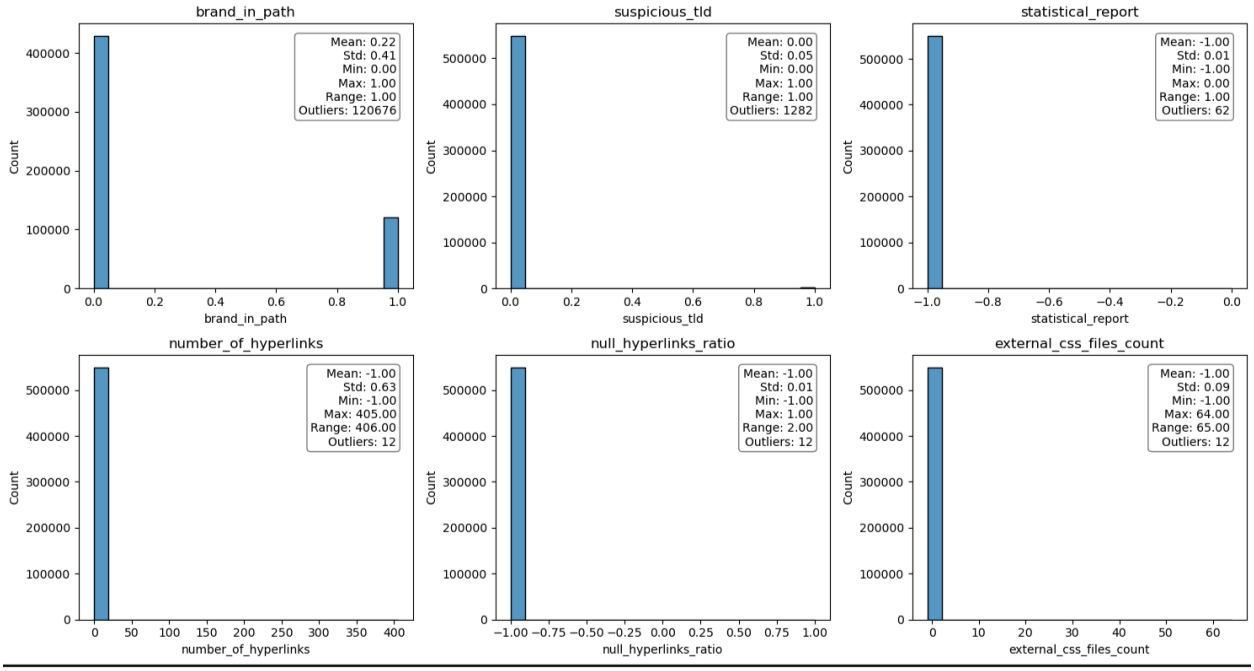

punycode_usage
Mean: 0.00
Std: 0.02
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 140
Count
punycode_usage
port_number_presence
Mean: 0.01
Std: 0.08
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 3137
Count
port_number_presence
tld_in_path
Mean: 0.99
Std: 0.09
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 4918
Count
tld_in_path
tld_in_subdomain
Mean: 0.00
Std: 0.01
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 54
Count
tld_in_subdomain
abnormal_subdomains
Mean: 0.05
Std: 0.22
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 26875
Count
abnormal_subdomains
number_of_subdomains
Mean: 0.52
Std: 1.05
Min: 0.00
Max: 33.00
Range: 33.00
Outliers: 9894
Count
number_of_subdomains

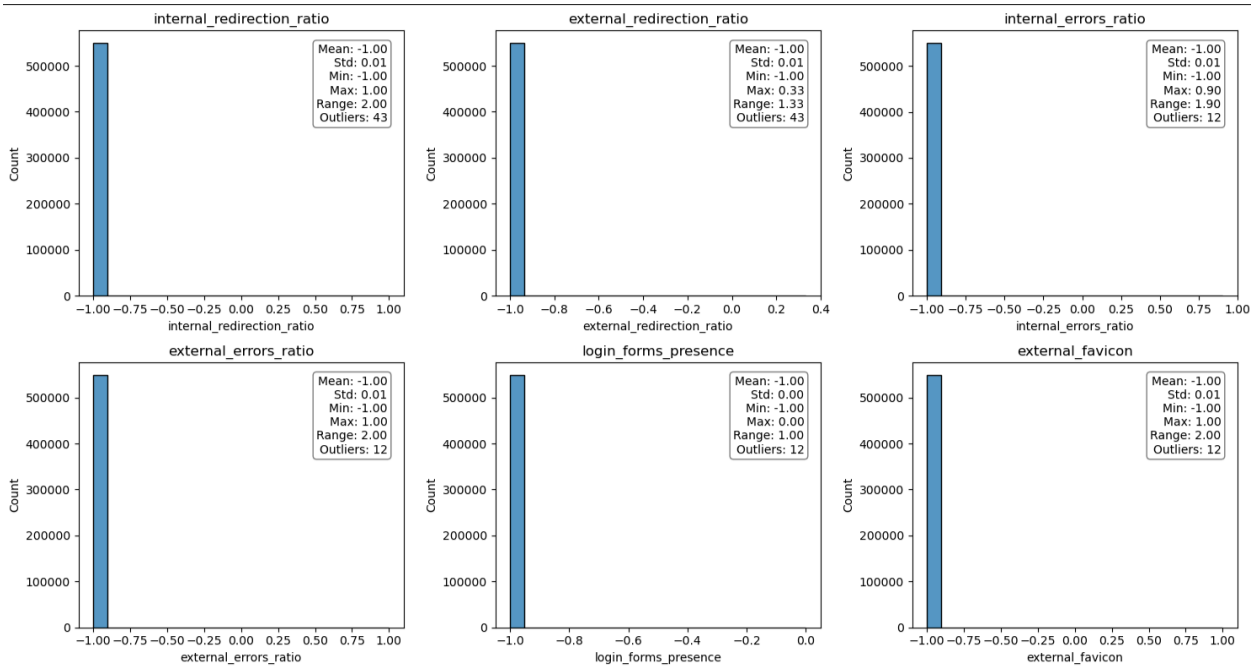

prefix_suffix_hyphen
Mean: 0.07
Std: 0.26
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 40744
Count
prefix_suffix_hyphen
random_domain_indicator
Mean: 0.26
Std: 0.44
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 0
Count
random_domain_indicator
url_shortening_service
Mean: 0.06
Std: 0.23
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 31384
Count
url_shortening_service
path_extension_check
Mean: 0.01
Std: 0.08
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 3180
Count
path_extension_check
redirection_count
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 3.00
Range: 4.00
Outliers: 42
Count
redirection_count
external_redirection_count
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 42
Count
external_redirection_count

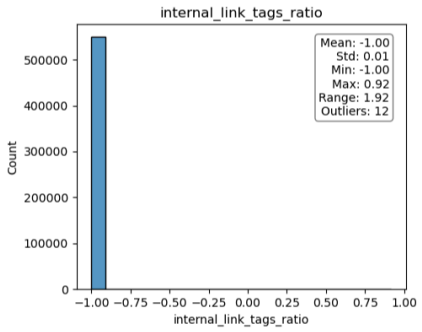

word_count_url
Mean: 7.11
Std: 5.15
Min: 0.00
Max: 166.00
Range: 166.00
Outliers: 36043
Count
word_count_url
word_count_hostname
Mean: 0.00
Std: 0.06
Min: 0.00
Max: 7.00
Range: 7.00
Outliers: 129
Count
word_count_hostname
word_count_path
Mean: 6.23
Std: 3.53
Min: 0.00
Max: 137.00
Range: 137.00
Outliers: 41060
Count
word_count_path
char_repeat_url
Mean: 1.43
Std: 1.36
Min: 0.00
Max: 200.00
Range: 200.00
Outliers: 964
Count
char_repeat_url
char_repeat_hostname
Mean: 0.00
Std: 0.03
Min: 0.00
Max: 3.00
Range: 3.00
Outliers: 60
Count
char_repeat_hostname
char_repeat_path
Mean: 1.34
Std: 1.17
Min: 0.00
Max: 29.00
Range: 29.00
Outliers: 482
Count
char_repeat_path

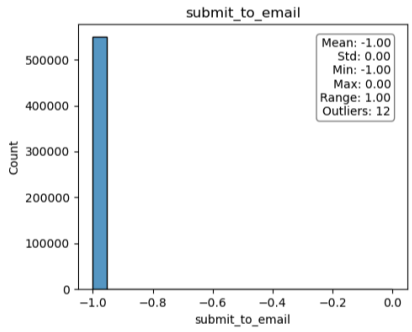

shortest_word_url
Mean: 2.35
Std: 0.74
Min: 0.00
Max: 19.00
Range: 19.00
Outliers: 279
shortest_word_hostname
Mean: 0.00
Std: 0.04
Min: 0.00
Max: 4.00
Range: 4.00
Outliers: 129
shortest_word_path
Mean: 2.47
Std: 0.68
Min: 0.00
Max: 19.00
Range: 19.00
Outliers: 390
longest_word_url
Mean: 14.40
Std: 15.01
Min: 0.00
Max: 2245.00
Range: 2245.00
Outliers: 43388
longest_word_hostname
Mean: 0.00
Std: 0.18
Min: 0.00
Max: 37.00
Range: 37.00
Outliers: 129
longest_word_path
Mean: 12.88
Std: 8.31
Min: 0.00
Max: 402.00
Range: 402.00
Outliers: 31335
Count

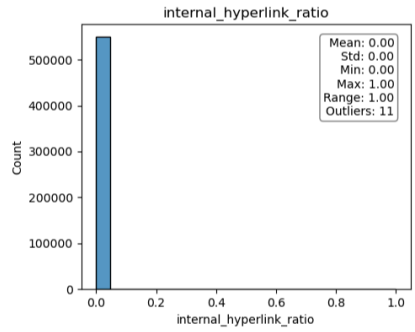

average_word_length_url
Mean: 6.37
Std: 2.56
Min: 0.00
Max: 328.71
Range: 328.71
Outliers: 31198
average_word_length_hostname
Mean: 0.00
Std: 0.09
Min: 0.00
Max: 14.67
Range: 14.67
Outliers: 129
average_word_length_path
Mean: 6.28
Std: 2.24
Min: 0.00
Max: 106.25
Range: 106.25
Outliers: 28800
phish_hints
Mean: 0.22
Std: 0.67
Min: 0.00
Max: 13.00
Range: 13.00
Outliers: 75157
brand_in_domain
Mean: 0.10
Std: 0.30
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 56519
brand_in_subdomain
Mean: 0.00
Std: 0.00
Min: 0.00
Max: 0.00
Range: 0.00
Outliers: 0
Count

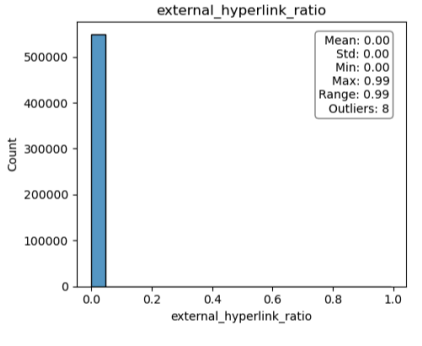

brand_in_path
Mean: 0.22
Std: 0.41
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 120676
Count
brand_in_path
suspicious_tld
Mean: 0.00
Std: 0.05
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 1282
Count
suspicious_tld
statistical_report
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 0.00
Range: 1.00
Outliers: 62
Count
statistical_report
number_of_hyperlinks
Mean: -1.00
Std: 0.63
Min: -1.00
Max: 405.00
Range: 406.00
Outliers: 12
Count
number_of_hyperlinks
null_hyperlinks_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
null_hyperlinks_ratio
external_css_files_count
Mean: -1.00
Std: 0.09
Min: -1.00
Max: 64.00
Range: 65.00
Outliers: 12
Count
external_css_files_count

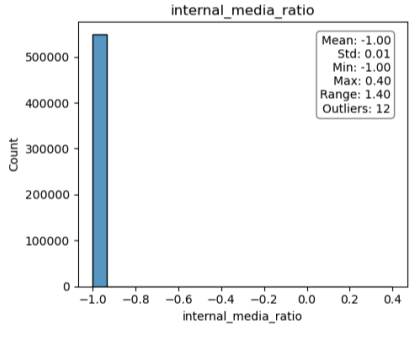

internal_redirection_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 43
Count
internal_redirection_ratio
external_redirection_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 0.33
Range: 1.33
Outliers: 43
Count
external_redirection_ratio
internal_errors_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 0.90
Range: 1.90
Outliers: 12
Count
internal_errors_ratio
external_errors_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
external_errors_ratio
login_forms_presence
Mean: -1.00
Std: 0.00
Min: -1.00
Max: 0.00
Range: 1.00
Outliers: 12
Count
login_forms_presence
external_favicon
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
external_favicon

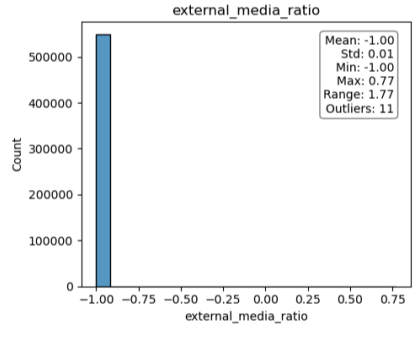

internal_link_tags_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 0.92
Range: 1.92
Outliers: 12
Count
internal_link_tags_ratio

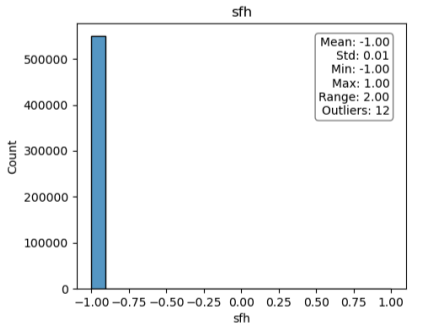

submit_to_email
Mean: -1.00
Std: 0.00
Min: -1.00
Max: 0.00
Range: 1.00
Outliers: 12
Count
submit_to_email

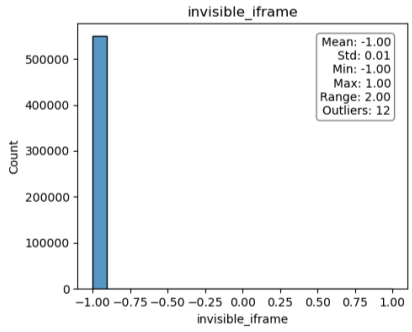

internal_hyperlink_ratio
Mean: 0.00
Std: 0.00
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 11
Count
internal_hyperlink_ratio

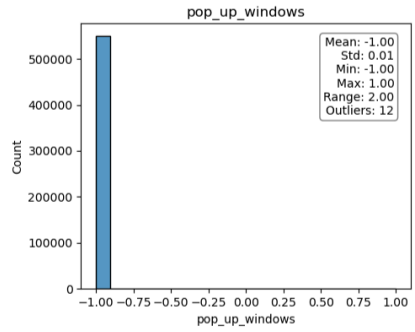

external_hyperlink_ratio
Mean: 0.00
Std: 0.00
Min: 0.00
Max: 0.99
Range: 0.99
Outliers: 8
Count
external_hyperlink_ratio

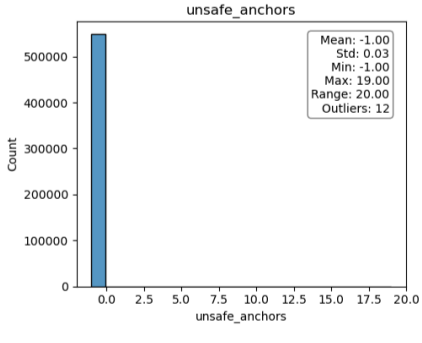

internal_media_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 0.40
Range: 1.40
Outliers: 12
Count
internal_media_ratio

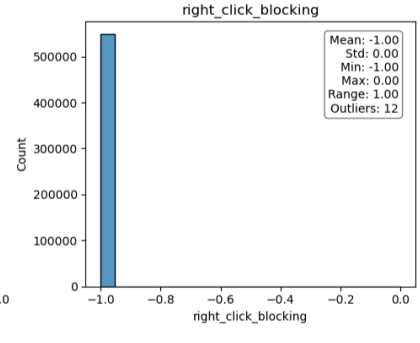

external_media_ratio
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 0.77
Range: 1.77
Outliers: 11
Count
external_media_ratio

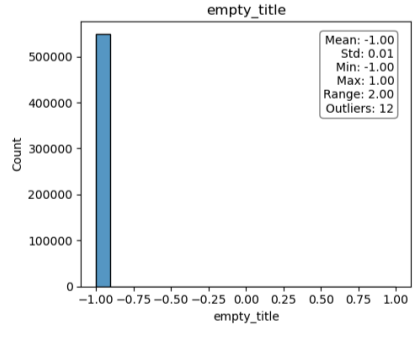

sfh
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
sfh

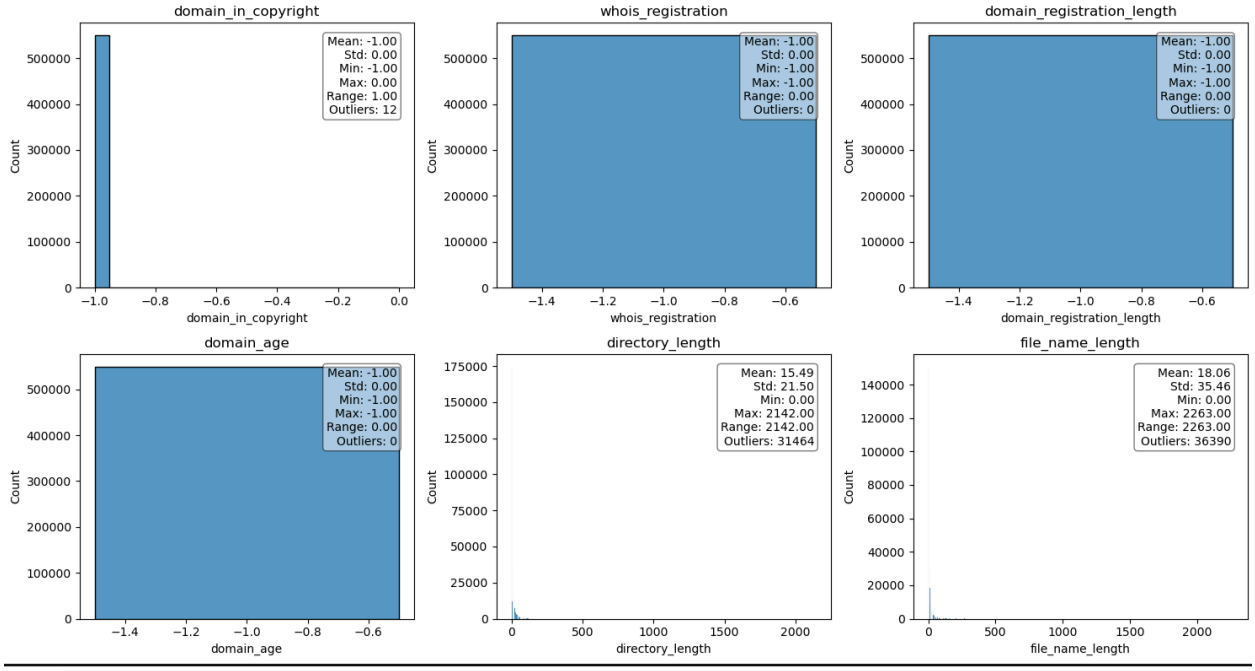

invisible_iframe
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
invisible_iframe

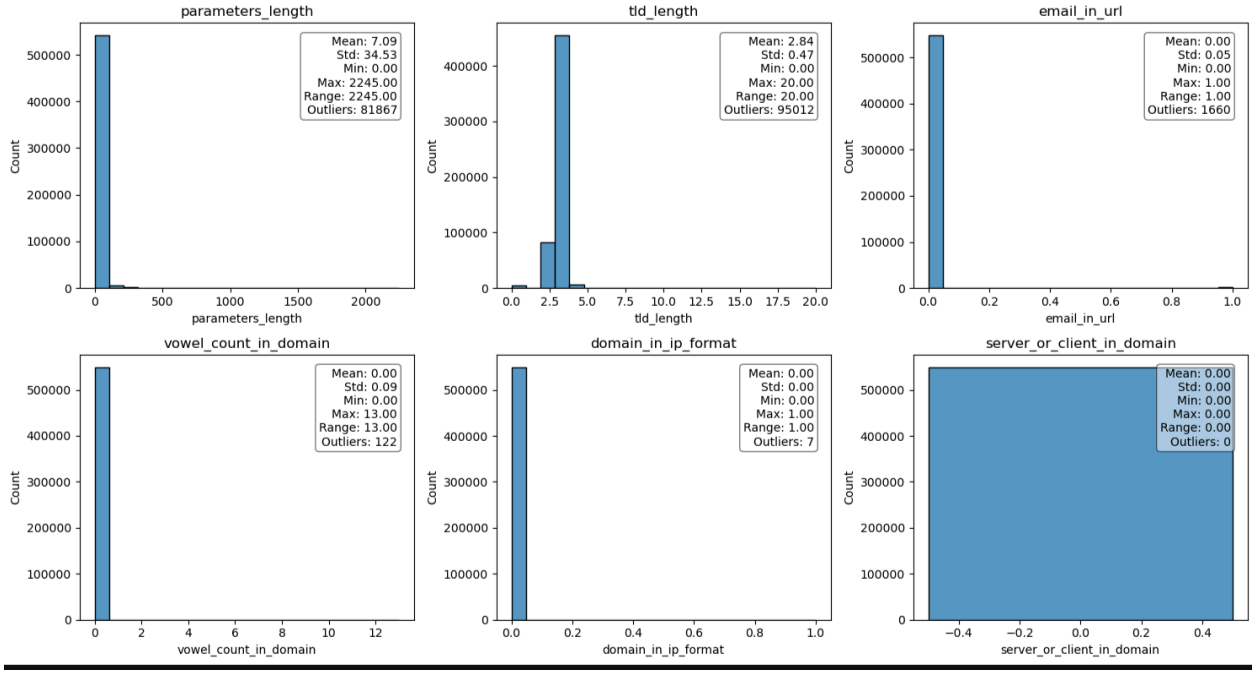

pop_up_windows
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
pop_up_windows

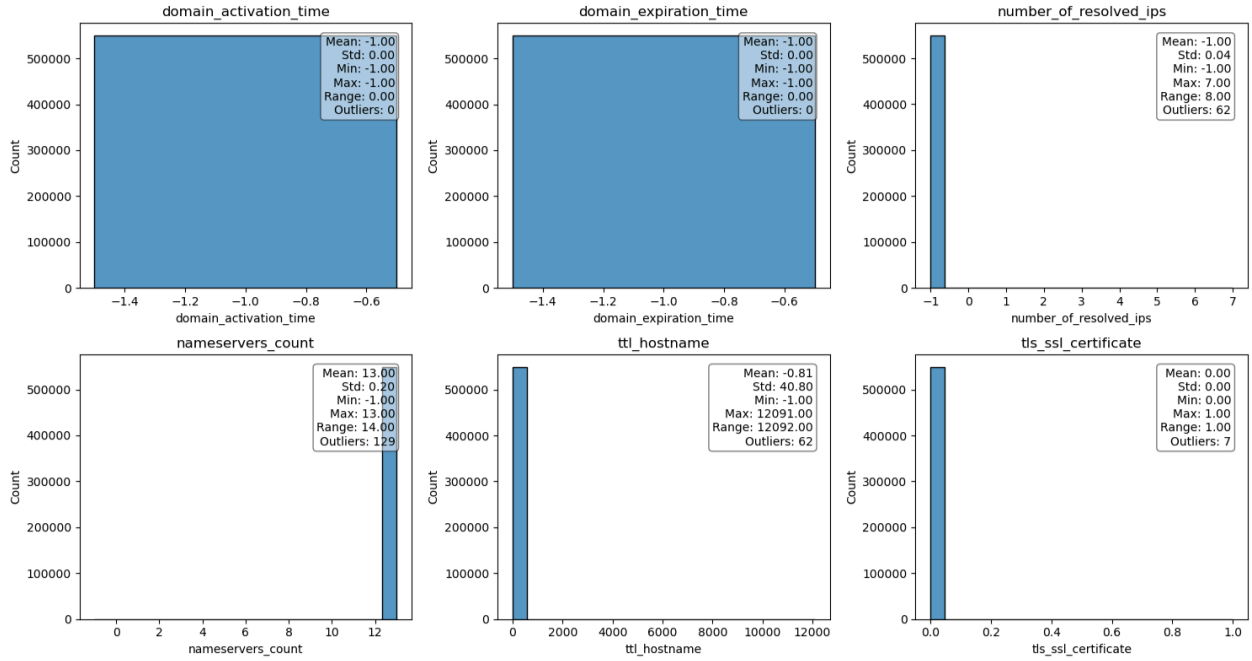

unsafe_anchors
Mean: -1.00
Std: 0.03
Min: -1.00
Max: 19.00
Range: 20.00
Outliers: 12
Count
unsafe_anchors

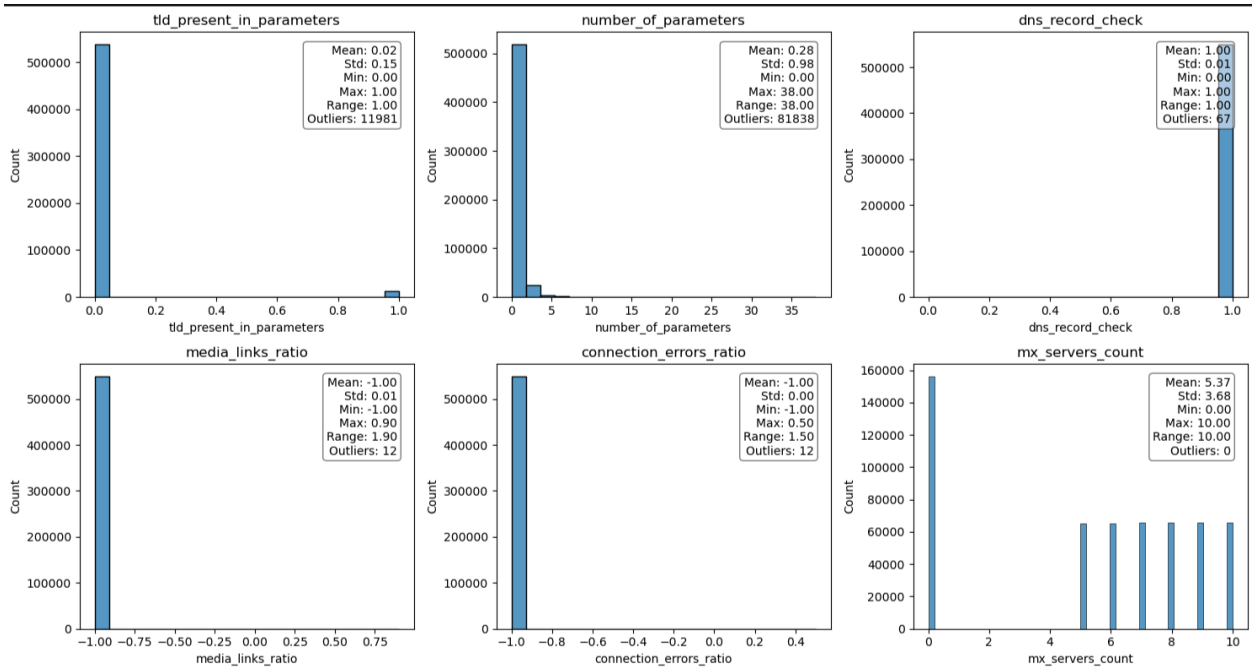

right_click_blocking
Mean: -1.00
Std: 0.00
Min: -1.00
Max: 0.00
Range: 1.00
Outliers: 12
Count
right_click_blocking


empty_title
Mean: -1.00
Std: 0.01
Min: -1.00
Max: 1.00
Range: 2.00
Outliers: 12
Count
empty_title


domain_in_copyright
Mean: -1.00
Std: 0.00
Min: -1.00
Max: 0.00
Range: 1.00
Outliers: 12
Count
domain_in_copyright
whois_registration
Mean: -1.00
Std: 0.00
Min: -1.00
Max: -1.00
Range: 0.00
Outliers: 0
Count
whois_registration
domain_registration_length
Mean: -1.00
Std: 0.00
Min: -1.00
Max: -1.00
Range: 0.00
Outliers: 0
Count
domain_registration_length
domain_age
Mean: -1.00
Std: 0.00
Min: -1.00
Max: -1.00
Range: 0.00
Outliers: 0
Count
domain_age
directory_length
Mean: 15.49
Std: 21.50
Min: 0.00
Max: 2142.00
Range: 2142.00
Outliers: 31464
Count
directory_length
file_name_length
Mean: 18.06
Std: 35.46
Min: 0.00
Max: 2263.00
Range: 2263.00
Outliers: 36390
Count
file_name_length

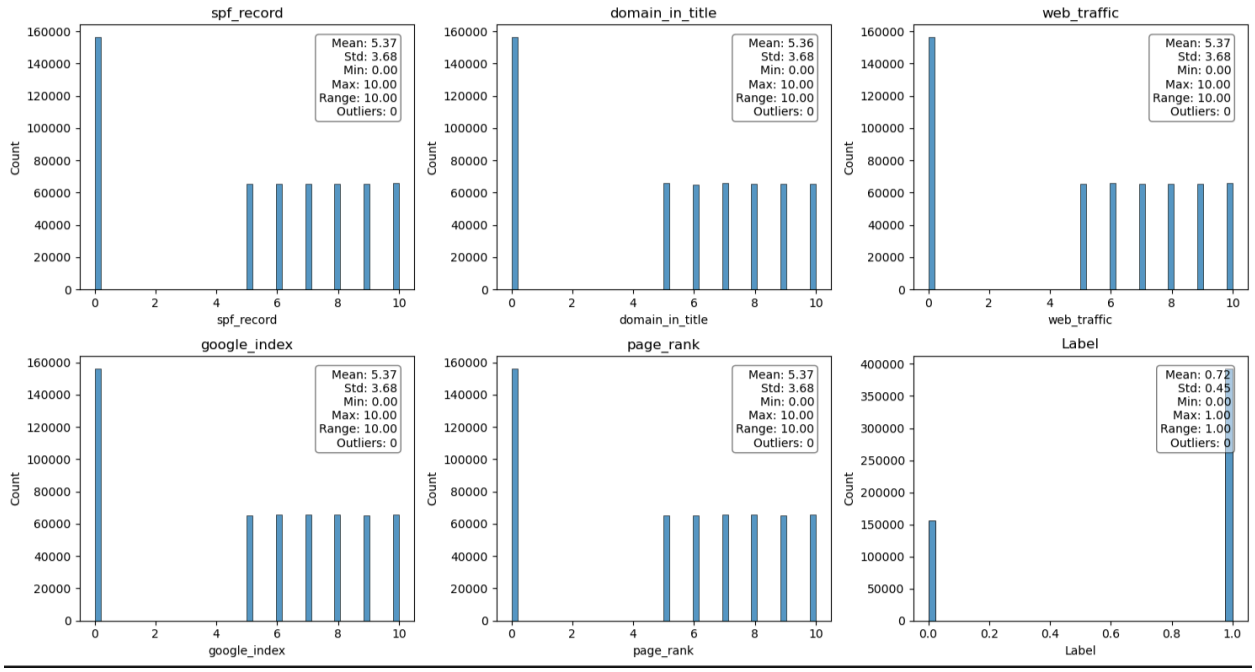

parameters_length
Mean: 7.09
Std: 34.53
Min: 0.00
Max: 2245.00
Range: 2245.00
Outliers: 81867
Count
parameters_length
tld_length
Mean: 2.84
Std: 0.47
Min: 0.00
Max: 20.00
Range: 20.00
Outliers: 95012
Count
tld_length
email_in_url
Mean: 0.00
Std: 0.05
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 1660
Count
email_in_url
vowel_count_in_domain
Mean: 0.00
Std: 0.09
Min: 0.00
Max: 13.00
Range: 13.00
Outliers: 122
Count
vowel_count_in_domain
domain_in_ip_format
Mean: 0.00
Std: 0.00
Min: 0.00
Max: 1.00
Range: 1.00
Outliers: 7
Count
domain_in_ip_format
server_or_client_in_domain
Mean: 0.00
Std: 0.00
Min: 0.00
Max: 0.00
Range: 0.00
Outliers: 0
Count
server_or_client_in_domain





We will z score normalize only those features which dont have gaussian distribution and have a huge range.

We will drop duplicate columns and highly correlated features.



Drop Repeated Columns

```
Column Index: 14, Column Name: tilde_count
Column Index: 28, Column Name: https_token
Column Index: 60, Column Name: brand_in_subdomain
Column Index: 86, Column Name: whois_registration
Column Index: 87, Column Name: domain_registration_length
Column Index: 88, Column Name: domain_age
Column Index: 96, Column Name: server_or_client_in_domain
Column Index: 98, Column Name: asn
Column Index: 99, Column Name: domain_activation_time
Column Index: 100, Column Name: domain_expiration_time
```

Drop highly correlated Columns with corr > 0.9

connection_errors_ratio,
'internal_link_tags_ratio',
'sfh',
'Nameservers_count',
 'Pop_up_windows',
 'internal_redirection_ratio',
'External_favicon',
 'Internal_media_ratio',
 'External_errors_ratio',
 'External_redirection_count',
 'dns_record_check',
'right_click_blocking',
'External_redirection_ratio',
'internal_errors_ratio',
'Domain_in_copyright',
 'Average_word_length_hostname',
 'number_of_parameters',
'Vowel_count_in_domain',
 'unsafe_anchors',
'Media_links_ratio',
 'login_forms_presence',
 'Empty_title',
 'Invisible_iframe',
 'Submit_to_email',
 'longest_word_hostname',
'external_media_ratio'