In [27]:
```
# Importing Data Manipilation Libraries
import pandas as pd
import numpy as np

# Import Data Visualization Libraries

import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# Import Data Filter Libraries
import warnings
warnings.filterwarnings('ignore')

# Import Data Logging Libraries
import logging
logging.basicConfig(level = logging.INFO,
                        filename = 'model.log',
                        filemode = 'w',
                        format = '%(asctime)s - %(levelname)s - %(message)s')

# Multicolinearity test and treatment libraries
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.decomposition import PCA
```

In [28]:
```
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', 100)
```

Loading Dataset

In [29]:
```
# Loading the dataset

url = 'https://raw.githubusercontent.com/mukeshmagar543/CODEB_Internship/refs/heads/main/dataset_ph

df = pd.read_csv(url)

df.sample(frac = 1) # Data Shuffle
```

Out[29]:

	url	length_url	length_hostname	ip	nb_dots	nb_hyphens
7993	http://www.asdnyi.com/house-mouse-facts/	40	14	0	2	2
6869	https://wiki.ezvid.com/best-wifi-radios	39	14	0	2	2
1453	http://pudhari.news	19	12	0	1	0
4859	http://www.mediacollege.com/video/shots/closeu...	52	20	0	3	0
9908	https://www.jogosonlinedemenina.com.br/	39	30	0	3	0
...	...	...	...	...	...	...
3529	http://www.instructables.com/id/Arduino-contro...	63	21	0	2	3
7303	http://usbank-link-mupyndtfft---com.illmickels...	105	50	1	3	5
3571	https://re-redirection-pp-account-id98763432.b...	58	49	1	2	4
2377	https://thecdm.ca/news/faculty-news/2013/10/15...	68	9	0	1	4
760	http://623112j4j3.codesandbox.io/kaifa	38	25	0	2	0

11430 rows × 89 columns

Getting Information about Dataset Like which column is object and which column is numerical

In [30]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11430 entries, 0 to 11429
Data columns (total 89 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   url                         11430 non-null  object 
 1   length_url                  11430 non-null  int64  
 2   length_hostname             11430 non-null  int64  
 3   ip                          11430 non-null  int64  
 4   nb_dots                     11430 non-null  int64  
 5   nb_hyphens                  11430 non-null  int64  
 6   nb_at                       11430 non-null  int64  
 7   nb_qm                       11430 non-null  int64  
 8   nb_and                      11430 non-null  int64  
 9   nb_or                       11430 non-null  int64  
 10  nb_eq                       11430 non-null  int64  
 11  nb_underscore               11430 non-null  int64  
 12  nb_tilde                    11430 non-null  int64  
 13  nb_percent                  11430 non-null  int64  
 14  nb_slash                    11430 non-null  int64  
 15  nb_star                     11430 non-null  int64  
 16  nb_colon                    11430 non-null  int64  
 17  nb_comma                    11430 non-null  int64  
 18  nb_semicolumn               11430 non-null  int64  
 19  nb_dollar                   11430 non-null  int64  
 20  nb_space                    11430 non-null  int64  
 21  nb_www                      11430 non-null  int64  
 22  nb_com                      11430 non-null  int64  
 23  nb_dslash                   11430 non-null  int64  
 24  http_in_path                11430 non-null  int64  
 25  https_token                 11430 non-null  int64  
 26  ratio_digits_url            11430 non-null  float64
 27  ratio_digits_host           11430 non-null  float64
 28  punycode                    11430 non-null  int64  
 29  port                        11430 non-null  int64  
 30  tld_in_path                 11430 non-null  int64  
 31  tld_in_subdomain            11430 non-null  int64  
 32  abnormal_subdomain          11430 non-null  int64  
 33  nb_subdomains               11430 non-null  int64  
 34  prefix_suffix               11430 non-null  int64  
 35  random_domain               11430 non-null  int64  
 36  shortening_service          11430 non-null  int64  
 37  path_extension              11430 non-null  int64  
 38  nb_redirection              11430 non-null  int64  
 39  nb_external_redirection     11430 non-null  int64  
 40  length_words_raw            11430 non-null  int64  
 41  char_repeat                 11430 non-null  int64  
 42  shortest_words_raw          11430 non-null  int64  
 43  shortest_word_host          11430 non-null  int64  
 44  shortest_word_path          11430 non-null  int64  
 45  longest_words_raw           11430 non-null  int64  
 46  longest_word_host           11430 non-null  int64  
 47  longest_word_path           11430 non-null  int64  
 48  avg_words_raw               11430 non-null  float64
 49  avg_word_host               11430 non-null  float64
 50  avg_word_path               11430 non-null  float64
 51  phish_hints                 11430 non-null  int64  
 52  domain_in_brand             11430 non-null  int64  
 53  brand_in_subdomain          11430 non-null  int64  
 54  brand_in_path               11430 non-null  int64  
 55  suspecious_tld              11430 non-null  int64  
 56  statistical_report          11430 non-null  int64  
 57  nb_hyperlinks               11430 non-null  int64  
 58  ratio_intHyperlinks         11430 non-null  float64
 59  ratio_extHyperlinks         11430 non-null  float64
 60  ratio_nullHyperlinks        11430 non-null  int64  
 61  nb_extCSS                   11430 non-null  int64  
 62  ratio_intRedirection        11430 non-null  int64  
 63  ratio_extRedirection        11430 non-null  float64
 64  ratio_intErrors             11430 non-null  int64  
 65  ratio_extErrors             11430 non-null  float64
 66  login_form                  11430 non-null  int64  
 67  external_favicon            11430 non-null  int64  
 68  links_in_tags               11430 non-null  float64
 69  submit_email                11430 non-null  int64
```

```
 70  ratio_intMedia              11430 non-null  float64
 71  ratio_extMedia              11430 non-null  float64
 72  sfh                         11430 non-null  int64
 73  iframe                      11430 non-null  int64
 74  popup_window                11430 non-null  int64
 75  safe_anchor                 11430 non-null  float64
 76  onmouseover                 11430 non-null  int64
 77  right_clic                  11430 non-null  int64
 78  empty_title                 11430 non-null  int64
 79  domain_in_title             11430 non-null  int64
 80  domain_with_copyright       11430 non-null  int64
 81  whois_registered_domain     11430 non-null  int64
 82  domain_registration_length  11430 non-null  int64
 83  domain_age                  11430 non-null  int64
 84  web_traffic                 11430 non-null  int64
 85  dns_record                  11430 non-null  int64
 86  google_index                11430 non-null  int64
 87  page_rank                   11430 non-null  int64
 88  status                      11430 non-null  object
dtypes: float64(13), int64(74), object(2)
memory usage: 7.8+ MB
```

Checking Null Values

- There is No Null Values are present in the given dataset.

In [31]:
```
df.isnull().sum()
```

Out[31]:
```
url                         0
length_url                  0
length_hostname             0
ip                          0
nb_dots                     0
nb_hyphens                  0
nb_at                       0
nb_qm                       0
nb_and                      0
nb_or                       0
nb_eq                       0
nb_underscore               0
nb_tilde                    0
nb_percent                  0
nb_slash                    0
nb_star                     0
nb_colon                    0
nb_comma                    0
nb_semicolumn               0
nb_dollar                   0
nb_space                    0
nb_www                      0
nb_com                      0
nb_dslash                   0
http_in_path                0
https_token                 0
ratio_digits_url            0
ratio_digits_host           0
punycode                    0
port                        0
tld_in_path                 0
tld_in_subdomain            0
abnormal_subdomain          0
nb_subdomains               0
prefix_suffix               0
random_domain               0
shortening_service          0
path_extension              0
nb_redirection              0
nb_external_redirection     0
length_words_raw            0
char_repeat                 0
shortest_words_raw          0
shortest_word_host          0
shortest_word_path          0
longest_words_raw           0
```

```
longest_word_host          0
longest_word_path          0
avg_words_raw              0
avg_word_host              0
avg_word_path              0
phish_hints                0
domain_in_brand            0
brand_in_subdomain         0
brand_in_path              0
suspecious_tld             0
statistical_report         0
nb_hyperlinks              0
ratio_intHyperlinks        0
ratio_extHyperlinks        0
ratio_nullHyperlinks       0
nb_extCSS                  0
ratio_intRedirection       0
ratio_extRedirection       0
ratio_intErrors            0
ratio_extErrors            0
login_form                 0
external_favicon           0
links_in_tags              0
submit_email               0
ratio_intMedia             0
ratio_extMedia             0
sfh                        0
iframe                     0
popup_window               0
safe_anchor                0
onmouseover                0
right_clic                 0
empty_title                0
domain_in_title            0
domain_with_copyright      0
whois_registered_domain    0
domain_registration_length 0
domain_age                 0
web_traffic                0
dns_record                 0
google_index               0
page_rank                  0
status                     0
dtype: int64
```

Descriptive Analysis

In [32]:
```
df.describe()
```

Out[32]:

	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm
count	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000
mean	61.126684	21.090289	0.150569	2.480752	0.997550	0.022222	0.141207
std	55.297318	10.777171	0.357644	1.369686	2.087087	0.155500	0.364456
min	12.000000	4.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	33.000000	15.000000	0.000000	2.000000	0.000000	0.000000	0.000000
50%	47.000000	19.000000	0.000000	2.000000	0.000000	0.000000	0.000000
75%	71.000000	24.000000	0.000000	3.000000	1.000000	0.000000	0.000000
max	1641.000000	214.000000	1.000000	24.000000	43.000000	4.000000	3.000000

Separating numerical and categorical columns. Then, for each numeric feature, you analyze spread, skewness, and outliers — very helpful for choosing scaling techniques or detecting which features might need transformation.

In [33]:
```
numerical_columns = df.select_dtypes(exclude= 'object')
numerical_columns
```

Out[33]:

	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_and	nb_or	nb_eq	nb_und
0	37	19	0	3	0	0	0	0	0	0	
1	77	23	1	1	0	0	0	0	0	0	
2	126	50	1	4	1	0	1	2	0	3	
3	18	11	0	2	0	0	0	0	0	0	
4	55	15	0	2	2	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	
11425	45	17	0	2	0	0	0	0	0	0	
11426	84	18	0	5	0	1	1	0	0	1	
11427	105	16	1	2	6	0	1	0	0	1	
11428	38	30	0	2	0	0	0	0	0	0	
11429	477	14	1	24	0	1	1	9	0	9	

11430 rows × 87 columns

In [34]:

```
# Descriptive statistics
from collections import OrderedDict

stats = []

for col in df.columns:
    if df[col].dtype != 'object':
        numerical_stats = OrderedDict({
            'Feature': col,
            'Minimum': df[col].min(),
            'Maximum': df[col].max(),
            'Mean': df[col].mean(),
            'Mode': df[col].mode()[0] if not df[col].mode().empty else None,
            '25%': df[col].quantile(0.25),
            '75%': df[col].quantile(0.75),
            'IQR': df[col].quantile(0.75) - df[col].quantile(0.25),
            'Standard Deviation': df[col].std(),
            'Skewness': df[col].skew(),
            'Kurtosis': df[col].kurt()
        })
        stats.append(numerical_stats)

# Convert to DataFrame
report = pd.DataFrame(stats)

report
```

Out[34]:

	Feature	Minimum	Maximum	Mean	Mode	25%	75%
0	length_url	12.0	1.641000e+03	61.126684	26.0	33.000000	71.000000
1	length_hostname	4.0	2.140000e+02	21.090289	16.0	15.000000	24.000000
2	ip	0.0	1.000000e+00	0.150569	0.0	0.000000	0.000000
3	nb_dots	1.0	2.400000e+01	2.480752	2.0	2.000000	3.000000
4	nb_hyphens	0.0	4.300000e+01	0.997550	0.0	0.000000	1.000000
5	nb_at	0.0	4.000000e+00	0.022222	0.0	0.000000	0.000000
6	nb_qm	0.0	3.000000e+00	0.141207	0.0	0.000000	0.000000
7	nb_and	0.0	1.900000e+01	0.162292	0.0	0.000000	0.000000
8	nb_or	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000
9	nb_eq	0.0	1.900000e+01	0.293176	0.0	0.000000	0.000000

10	nb_underscore	0.0	1.800000e+01	0.322660	0.0	0.000000	0.000000
11	nb_tilde	0.0	1.000000e+00	0.006649	0.0	0.000000	0.000000
12	nb_percent	0.0	9.600000e+01	0.123097	0.0	0.000000	0.000000
13	nb_slash	2.0	3.300000e+01	4.289589	3.0	3.000000	5.000000
14	nb_star	0.0	1.000000e+00	0.000700	0.0	0.000000	0.000000
15	nb_colon	1.0	7.000000e+00	1.027909	1.0	1.000000	1.000000
16	nb_comma	0.0	4.000000e+00	0.004024	0.0	0.000000	0.000000
17	nb_semicolumn	0.0	2.000000e+01	0.062292	0.0	0.000000	0.000000
18	nb_dollar	0.0	6.000000e+00	0.001925	0.0	0.000000	0.000000
19	nb_space	0.0	1.800000e+01	0.034821	0.0	0.000000	0.000000
20	nb_www	0.0	2.000000e+00	0.448469	0.0	0.000000	1.000000
21	nb_com	0.0	6.000000e+00	0.127997	0.0	0.000000	0.000000
22	nb_dslash	0.0	1.000000e+00	0.006562	0.0	0.000000	0.000000
23	http_in_path	0.0	4.000000e+00	0.016710	0.0	0.000000	0.000000
24	https_token	0.0	1.000000e+00	0.610936	1.0	0.000000	1.000000
25	ratio_digits_url	0.0	7.238806e-01	0.053137	0.0	0.000000	0.079365
26	ratio_digits_host	0.0	8.000000e-01	0.025024	0.0	0.000000	0.000000
27	punycode	0.0	1.000000e+00	0.000350	0.0	0.000000	0.000000
28	port	0.0	1.000000e+00	0.002362	0.0	0.000000	0.000000
29	tld_in_path	0.0	1.000000e+00	0.065617	0.0	0.000000	0.000000
30	tld_in_subdomain	0.0	1.000000e+00	0.050131	0.0	0.000000	0.000000
31	abnormal_subdomain	0.0	1.000000e+00	0.021610	0.0	0.000000	0.000000
32	nb_subdomains	1.0	3.000000e+00	2.231671	2.0	2.000000	3.000000
33	prefix_suffix	0.0	1.000000e+00	0.202450	0.0	0.000000	0.000000
34	random_domain	0.0	1.000000e+00	0.083290	0.0	0.000000	0.000000
35	shortening_service	0.0	1.000000e+00	0.123447	0.0	0.000000	0.000000
36	path_extension	0.0	1.000000e+00	0.000175	0.0	0.000000	0.000000
37	nb_redirection	0.0	6.000000e+00	0.498250	0.0	0.000000	1.000000
38	nb_external_redirection	0.0	1.000000e+00	0.003150	0.0	0.000000	0.000000
39	length_words_raw	1.0	1.060000e+02	6.232808	2.0	2.000000	8.000000
40	char_repeat	0.0	1.460000e+02	2.927472	3.0	1.000000	4.000000
41	shortest_words_raw	1.0	3.100000e+01	3.127297	3.0	2.000000	3.000000
42	shortest_word_host	1.0	3.900000e+01	5.019773	3.0	3.000000	6.000000
43	shortest_word_path	0.0	4.000000e+01	2.398950	0.0	0.000000	3.000000
44	longest_words_raw	2.0	8.290000e+02	15.393876	9.0	9.000000	16.000000
45	longest_word_host	1.0	6.200000e+01	10.467979	9.0	7.000000	13.000000
46	longest_word_path	0.0	8.290000e+02	10.561505	0.0	0.000000	11.000000
47	avg_words_raw	2.0	1.282500e+02	7.258882	6.0	5.250000	8.000000
48	avg_word_host	1.0	3.900000e+01	7.678075	5.0	5.250000	9.000000
49	avg_word_path	0.0	2.500000e+02	5.092425	0.0	0.000000	6.714286
50	phish_hints	0.0	1.000000e+01	0.327734	0.0	0.000000	0.000000
51	domain_in_brand	0.0	1.000000e+00	0.104199	0.0	0.000000	0.000000
52	brand_in_subdomain	0.0	1.000000e+00	0.004112	0.0	0.000000	0.000000

53	brand_in_path	0.0	1.000000e+00	0.004899	0.0	0.000000	0.000000	
54	suspecious_tld	0.0	1.000000e+00	0.017935	0.0	0.000000	0.000000	
55	statistical_report	0.0	2.000000e+00	0.059755	0.0	0.000000	0.000000	
56	nb_hyperlinks	0.0	4.659000e+03	87.189764	0.0	9.000000	101.000000	
57	ratio_intHyperlinks	0.0	1.000000e+00	0.602457	0.0	0.224991	0.944767	
58	ratio_extHyperlinks	0.0	1.000000e+00	0.276720	0.0	0.000000	0.474840	
59	ratio_nullHyperlinks	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	
60	nb_extCSS	0.0	1.240000e+02	0.784864	0.0	0.000000	1.000000	
61	ratio_intRedirection	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	
62	ratio_extRedirection	0.0	2.000000e+00	0.158926	0.0	0.000000	0.230769	
63	ratio_intErrors	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	
64	ratio_extErrors	0.0	1.000000e+00	0.062469	0.0	0.000000	0.034483	
65	login_form	0.0	1.000000e+00	0.063605	0.0	0.000000	0.000000	
66	external_favicon	0.0	1.000000e+00	0.442170	0.0	0.000000	1.000000	
67	links_in_tags	0.0	1.000000e+02	51.978211	0.0	0.000000	98.061004	
68	submit_email	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	
69	ratio_intMedia	0.0	1.000000e+02	42.870444	0.0	0.000000	100.000000	
70	ratio_extMedia	0.0	1.000000e+02	23.236293	0.0	0.000000	33.333333	
71	sfh	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	
72	iframe	0.0	1.000000e+00	0.001312	0.0	0.000000	0.000000	
73	popup_window	0.0	1.000000e+00	0.006037	0.0	0.000000	0.000000	
74	safe_anchor	0.0	1.000000e+02	37.063922	0.0	0.000000	75.000000	
75	onmouseover	0.0	1.000000e+00	0.001137	0.0	0.000000	0.000000	
76	right_clic	0.0	1.000000e+00	0.001400	0.0	0.000000	0.000000	
77	empty_title	0.0	1.000000e+00	0.124759	0.0	0.000000	0.000000	
78	domain_in_title	0.0	1.000000e+00	0.775853	1.0	1.000000	1.000000	
79	domain_with_copyright	0.0	1.000000e+00	0.439545	0.0	0.000000	1.000000	
80	whois_registered_domain	0.0	1.000000e+00	0.072878	0.0	0.000000	0.000000	
81	domain_registration_length	-1.0	2.982900e+04	492.532196	0.0	84.000000	449.000000	
82	domain_age	-12.0	1.287400e+04	4062.543745	-1.0	972.250000	7026.750000	6
83	web_traffic	0.0	1.076799e+07	856756.643307	0.0	0.000000	373845.500000	373
84	dns_record	0.0	1.000000e+00	0.020122	0.0	0.000000	0.000000	
85	google_index	0.0	1.000000e+00	0.533946	1.0	0.000000	1.000000	
86	page_rank	0.0	1.000000e+01	3.185739	0.0	1.000000	5.000000	

Frequency distribution for categorical features

Several features showed significant skewness, suggesting non-normal distributions.

Wide ranges and high standard deviations in some columns (e.g., web_traffic, length_url) indicate the presence of outliers.

Features with high kurtosis are likely to have heavy tails or sharp peaks.

Checking frequency counts for categorical columns — this helps you see whether categories are balanced or

dominated by one class (like the target label status).

In [35]:
```
# Frequency distribution for categorical features (if any)
for col in df.columns:
    if df[col].dtype == 'object':
        print(f"\nFrequency distribution for {col}:\n")
        print(df[col].value_counts())
```

```
Frequency distribution for url:

url
http://e710z0ear.du.r.appspot.com/c:/users/user/downlo
2
https://lt.mydplr.com/16672ac75448ecdb528e1c663c0df3a7-f10ed321df1a4fbc893c86fbb12f0913
1
http://appleid.apple.com-app.es/
1
http://174.139.46.123/ap/signin?openid.pape.max_auth_age=0&amp;openid.return_to=https%3A%2F%2Fwww.ama
zon.co.jp%2F%3Fref_%3Dnav_em_hd_re_signin&amp;openid.identity=http%3A%2F%2Fspecs.openid.net%2Fauth%2F
2.0%2Fidentifier_select&amp;openid.assoc_handle=jpflex&amp;openid.mode=checkid_setup&amp;key=a@b.c&am
p;openid.claimed_id=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0%2Fidentifier_select&amp;openid.ns=htt
p%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0&amp;&amp;ref_=nav_em_hd_clc_signin    1
http://www.crestonwood.com/router.php
1

..
https://www.dissernet.org/
1
https://workprotocoles-com.webs.com/
1
http://www.vg247.com/2017/04/24/best-nintendo-switch-games/
1
https://www.facebook.com/Publictransporthub/
1
http://www.game.co.uk/en/games/nintendo-switch/nintendo-switch/
1
Name: count, Length: 11429, dtype: int64

Frequency distribution for status:

status
legitimate    5715
phishing      5715
Name: count, dtype: int64
```

The target label is balanced — There is no need to use SMOTE techniques to Blanace the Target columm.

In [36]:
```
df['status'].mode()
```

Out[36]:
```
0    legitimate
1      phishing
Name: status, dtype: object
```

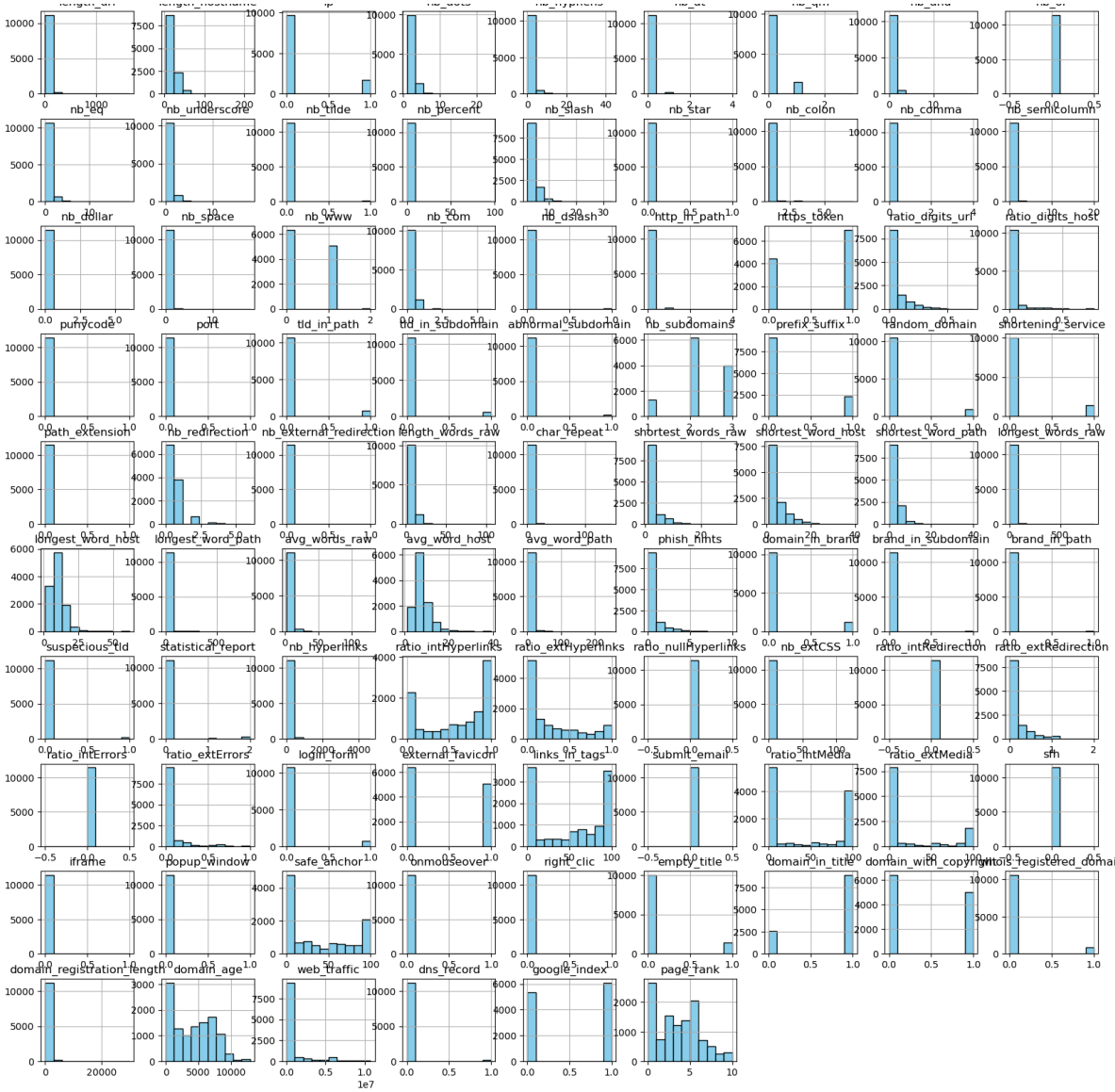
In [37]:
```
df['url'].mode()
```

Out[37]:
```
0    http://e710z0ear.du.r.appspot.com/c:/users/use...
Name: url, dtype: object
```

Histogram

Histograms Reveal skewed features and possible outliers. Some features like web_traffic or length_url may need scaling or normalization.

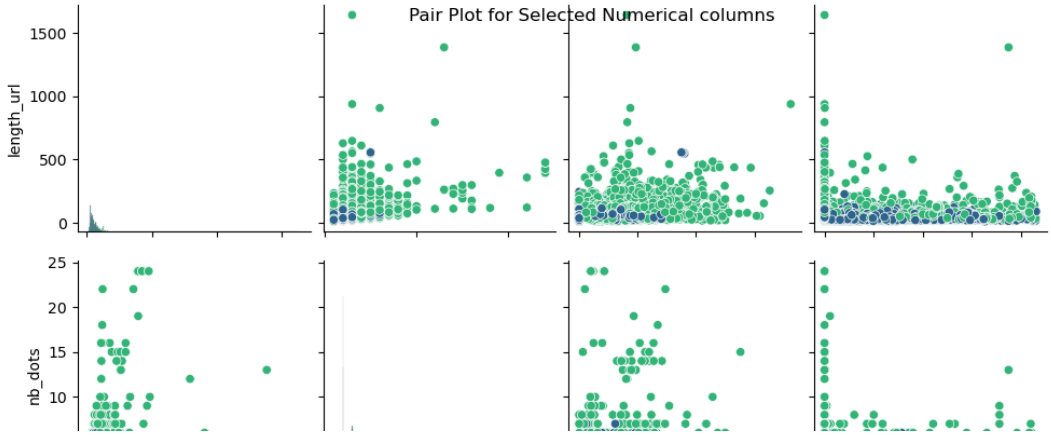
In [38]:
```
# Histograms for numerical features
numerical_columns.hist(figsize=(20, 20),bins= 10, color= 'skyblue', edgecolor= 'black')
plt.title("Histogram")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```

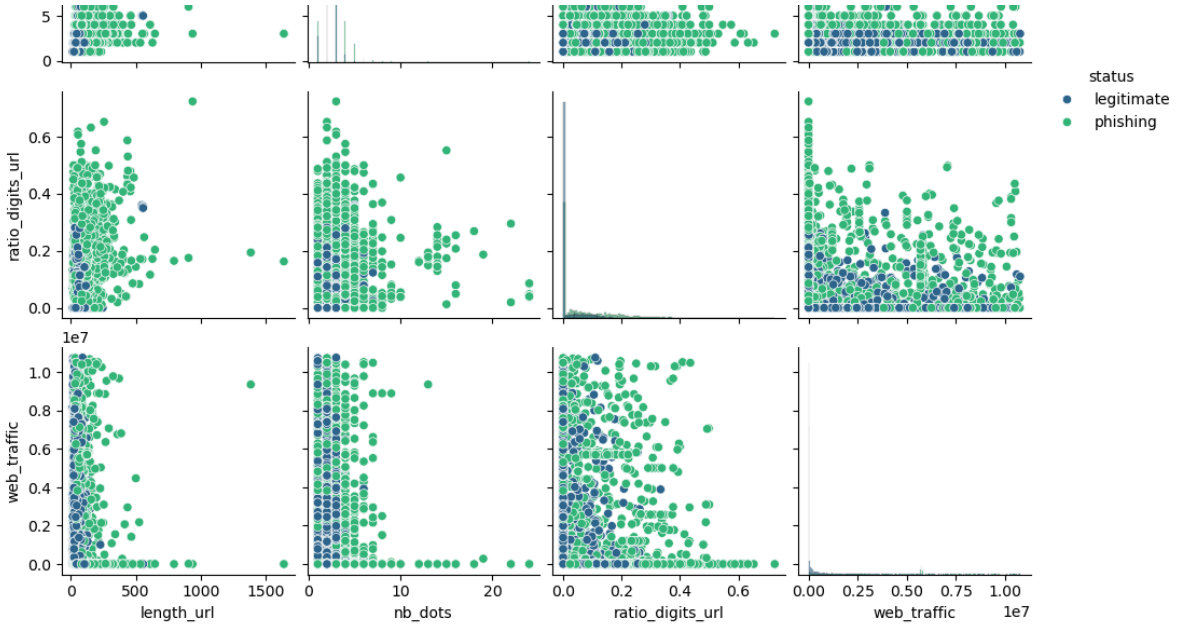


Pair Plot

- We have use only selected important features to create the Pair Plot
- The pairplot shows some visual separation between phishing and legitimate classes in selected features — especially in ratio_digits_url and web_traffic. That means these features might be strong indicators for classification.

In [39]:
```
selected_features = ['length_url', 'nb_dots', 'ratio_digits_url', 'web_traffic', 'status']
# plot pair plot
sns.pairplot(df[selected_features], hue='status', diag_kind='hist', palette= 'viridis')
plt.suptitle('Pair Plot for Selected Numerical columns')
plt.show()
```





Using Replace function to 'legitimate' and 'phishing' into 0 and 1 — readying the target for machine learning models.

In [40]:
```
df['status'] = df['status'].replace({'legitimate' : 0, 'phishing' : 1})
```

Label encoding to url column — to convert the categorical data into numerical

In [41]:
```
# Using Label Encoding in Url column
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

df['url'] = le.fit_transform(df['url'])
df['url'].value_counts()
```

Out[41]:
```
url
1065    2
8258    1
363     1
62      1
4501    1
       ..
9799    1
9324    1
6684    1
9920    1
4919    1
Name: count, Length: 11429, dtype: int64
```

Insights and Recommendations

- Features like `web_traffic`, `SSLfinal_State`, and `page_rank` are crucial indicators.
- The Dataset has huge amount of Outliers.
- Outliers can be capped using the IQR method.
- Use `RobustScaler` to normalize numerical features.
- Remove redundant features with high multicollinearity.
- The target is balance hence, there is no need for SMOTE.
- We can use Feature Engineering.
- The Dataset have doesn't have any null values.

Checking Duplicates

Label Encoding was applied to the url column to convert categorical values into numeric form. One-Hot

Encoding was avoided because it would have significantly increased the number of columns due to the high number of unique URLs. Label Encoding keeps the dataset compact and efficient without adding unnecessary dimensions.

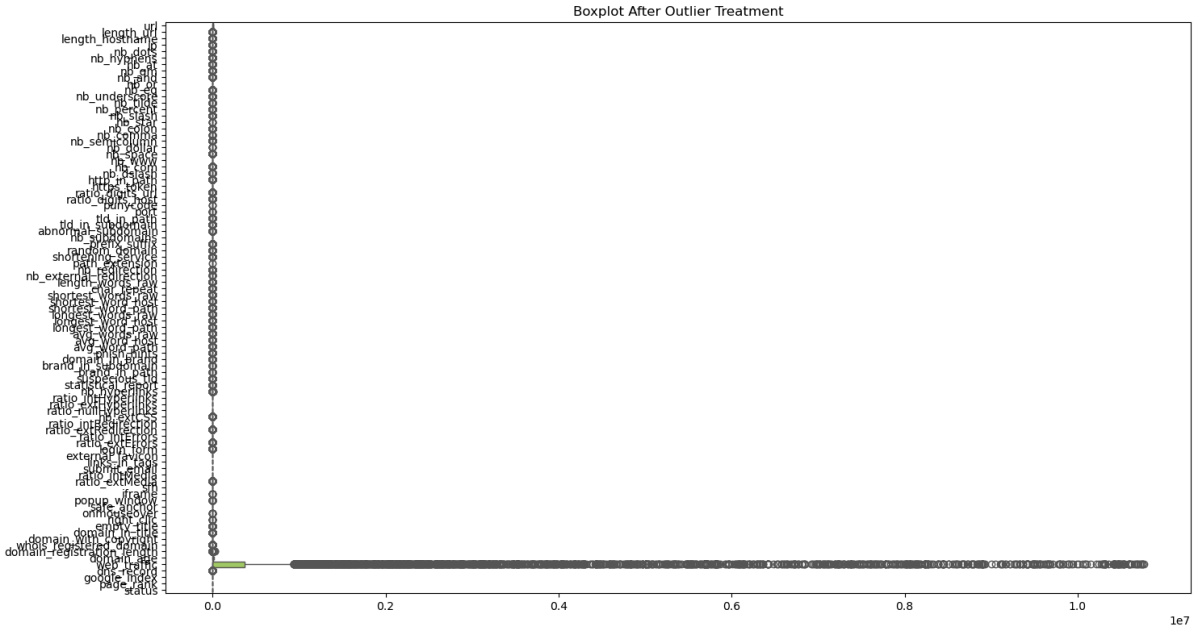
In [42]:
```
# Checking Duplicates
duplicates = df.duplicated()
duplicates.value_counts()
```

Out[42]:
```
False    11430
Name: count, dtype: int64
```

In [43]:
```
# Set figure size
plt.figure(figsize=(15, 8))

# Create boxplot for all numerical columns
sns.boxplot(data=df, orient='h', palette='Set2')

# Set title
plt.title('Boxplot After Outlier Treatment')
plt.tight_layout()
plt.show()
```



A ranked list of features based on Variance Variance Inflation Factor (VIF)

In [44]:
```
# Checking VIF:
def calculate_vif(dataset):
    vif = pd.DataFrame()
    vif['features'] = dataset.columns
    vif['VIF_Values'] = [variance_inflation_factor(dataset.values,i) for i in range(dataset.shape[1
    vif['VIF_Values'] = round(vif['VIF_Values'], 2)
    vif = vif.sort_values(by = 'VIF_Values', ascending=False)
    return (vif)

calculate_vif(df.drop('status',axis = 1))
```

Out[44]:

	features	VIF_Values
49	avg_word_host	278.79
45	longest_words_raw	150.19
40	length_words_raw	144.10
47	longest_word_path	130.30

46	longest_word_host	127.15
48	avg_words_raw	92.81
43	shortest_word_host	51.16
14	nb_slash	45.65
4	nb_dots	34.09
33	nb_subdomains	33.03
16	nb_colon	29.59
0	url	28.20
1	length_url	25.48
50	avg_word_path	25.29
58	ratio_intHyperlinks	21.28
2	length_hostname	19.04
10	nb_eq	14.34
25	https_token	14.33
8	nb_and	12.27
42	shortest_words_raw	11.80
5	nb_hyphens	11.15
68	links_in_tags	8.07
87	page_rank	7.48
59	ratio_extHyperlinks	7.34
21	nb_www	6.31
79	domain_in_title	5.99
13	nb_percent	5.14
83	domain_age	5.08
26	ratio_digits_url	4.94
7	nb_qm	4.17
86	google_index	4.16
11	nb_underscore	4.02
3	ip	4.01
44	shortest_word_path	3.86
27	ratio_digits_host	3.73
70	ratio_intMedia	3.73
67	external_favicon	3.29
78	empty_title	3.07
75	safe_anchor	3.00
31	tld_in_subdomain	2.74
24	http_in_path	2.59
71	ratio_extMedia	2.52
22	nb_com	2.38
41	char_repeat	2.29
80	domain_with_copyright	2.21
32	abnormal_subdomain	2.17
52	domain_in_brand	2.05

51	phish_hints	1.97
38	nb_redirection	1.95
30	tld_in_path	1.87
18	nb_semicolumn	1.86
34	prefix_suffix	1.78
57	nb_hyperlinks	1.71
85	dns_record	1.69
63	ratio_extRedirection	1.66
82	domain_registration_length	1.66
39	nb_external_redirection	1.64
36	shortening_service	1.57
23	nb_dslash	1.52
56	statistical_report	1.51
84	web_traffic	1.47
54	brand_in_path	1.37
65	ratio_extErrors	1.34
61	nb_extCSS	1.33
81	whois_registered_domain	1.32
6	nb_at	1.30
35	random_domain	1.20
66	login_form	1.16
20	nb_space	1.15
53	brand_in_subdomain	1.14
29	port	1.14
55	suspecious_tld	1.10
12	nb_tilde	1.06
19	nb_dollar	1.05
76	onmouseover	1.04
17	nb_comma	1.04
15	nb_star	1.03
28	punycode	1.02
74	popup_window	1.02
77	right_clic	1.01
73	iframe	1.01
37	path_extension	1.00
9	nb_or	NaN
60	ratio_nullHyperlinks	NaN
62	ratio_intRedirection	NaN
64	ratio_intErrors	NaN
69	submit_email	NaN
72	sfh	NaN

In [45]:
```
# Splitting Data into Independent And target Column
```

```
X=df.drop(columns='status')
y=df['status']
```

In [46]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,train_size=0.70,random_state=42)
```

In [47]:
```
X_train_original = X_train.copy()
```

Scaling Technique:- Robust Scaler

Robust Scaler was used to handle outliers effectively, as boxplots showed many extreme values in the numerical features. It scales data based on the median and IQR, making it less sensitive to outliers compared to StandardScaler or MinMaxScaler.

In [48]:
```
from sklearn.preprocessing import MinMaxScaler,StandardScaler,RobustScaler
scaler=RobustScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.transform(X_test)
```

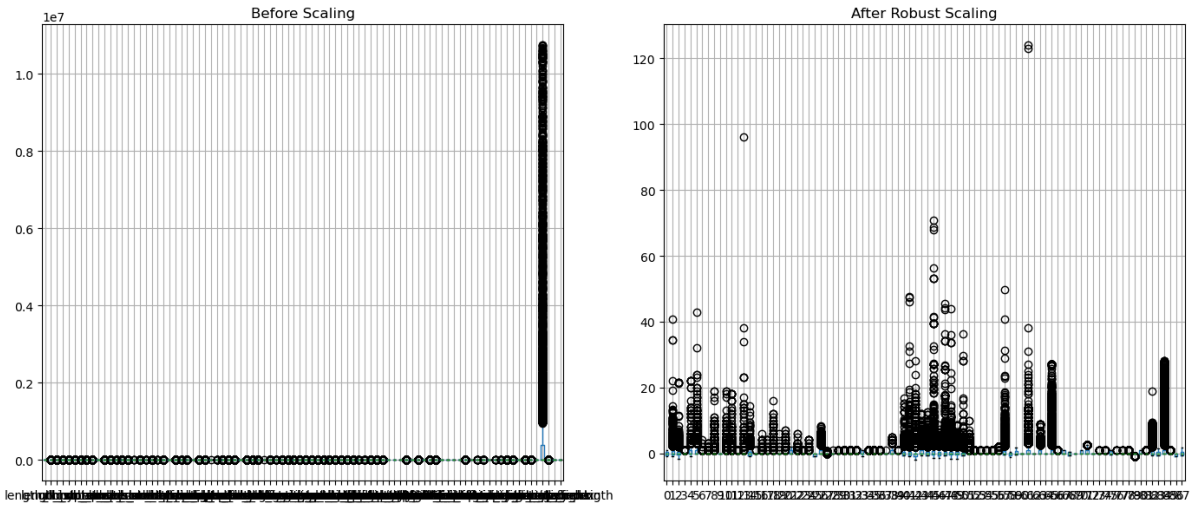
In [49]:
```
X_train_scaled=X_train.copy()
# If X_train is a NumPy array, convert it to a DataFrame
X_train_df = pd.DataFrame(X_train_original)
X_train_scaled_df = pd.DataFrame(X_train_scaled)

# Plot before and after scaling side by side
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
X_train_df.boxplot()
plt.title("Before Scaling")

plt.subplot(1, 2, 2)
X_train_scaled_df.boxplot()
plt.title("After Robust Scaling")

plt.tight_layout()
plt.show()
```



In [ ]:
```
# Table summarizing feature correlations
df.corr()['status']
```

Out[ ]:
```
url                          -2.909714e-01
length_url                    2.485805e-01
length_hostname               2.383224e-01
ip                            3.216978e-01
nb_dots                       2.070288e-01
nb_hyphens                   -1.001075e-01
nb_at                         1.429146e-01
```

```
nb_qm                        2.943191e-01
nb_and                       1.705464e-01
nb_or                                 NaN
nb_eq                        2.333863e-01
nb_underscore                3.809134e-02
nb_tilde                     3.014233e-02
nb_percent                   2.810129e-02
nb_slash                     2.422700e-01
nb_star                      2.646512e-02
nb_colon                     9.283531e-02
nb_comma                     1.186465e-02
nb_semicolumn                1.035541e-01
nb_dollar                    2.496206e-02
nb_space                    -4.193222e-03
nb_www                      -4.434677e-01
nb_com                       1.562835e-01
nb_dslash                    7.260234e-02
http_in_path                 7.077624e-02
https_token                  1.146691e-01
ratio_digits_url             3.563946e-01
ratio_digits_host            2.243349e-01
punycode                     1.871039e-02
port                         9.011116e-03
tld_in_path                  7.914651e-02
tld_in_subdomain             2.088842e-01
abnormal_subdomain           1.281598e-01
nb_subdomains                1.128907e-01
prefix_suffix                2.146807e-01
random_domain                1.963062e-02
shortening_service           1.061200e-01
path_extension               5.592660e-17
nb_redirection              -2.440520e-02
nb_external_redirection      5.620994e-02
length_words_raw             1.920105e-01
char_repeat                  1.473217e-02
shortest_words_raw          -3.936361e-02
shortest_word_host           2.230840e-01
shortest_word_path           7.436495e-02
longest_words_raw            2.001466e-01
longest_word_host            1.245156e-01
longest_word_path            2.127091e-01
avg_words_raw                1.675637e-01
avg_word_host                1.935017e-01
avg_word_path                1.972561e-01
phish_hints                  3.353927e-01
domain_in_brand             -9.822216e-02
brand_in_subdomain           6.425702e-02
brand_in_path                6.515575e-02
suspecious_tld               1.100896e-01
statistical_report           1.439435e-01
nb_hyperlinks               -3.426283e-01
ratio_intHyperlinks         -2.439821e-01
ratio_extHyperlinks          8.335725e-02
ratio_nullHyperlinks                  NaN
nb_extCSS                   -8.356663e-02
ratio_intRedirection                  NaN
ratio_extRedirection        -1.508267e-01
ratio_intErrors                       NaN
ratio_extErrors             -3.470251e-02
login_form                  -1.900010e-02
external_favicon            -1.465654e-01
links_in_tags               -1.844011e-01
submit_email                          NaN
ratio_intMedia              -1.933331e-01
ratio_extMedia              -1.404059e-01
sfh                                   NaN
iframe                      -1.208332e-02
popup_window                -5.760197e-02
safe_anchor                 -1.733973e-01
onmouseover                 -7.787061e-03
right_clic                   4.680056e-03
empty_title                  2.070428e-01
domain_in_title              3.428070e-01
domain_with_copyright       -1.730985e-01
whois_registered_domain      6.697907e-02
domain_registration_length  -1.617188e-01
domain_age                  -3.318891e-01
web_traffic                  6.038772e-02
dns_record                   1.231109e-01
```

```
dns_record                  1.221190e-01
google_index                7.311708e-01
page_rank                  -5.111371e-01
status                      1.000000e+00
Name: status, dtype: float64
```