CodeB_Internship / modell.ipynb

MontyVasita week 3 Done\ dd98edc · 5 days ago History

Preview Code Blame 4913 lines (4913 loc) · 1.33 MB Raw

Monty K Vasita

Setting up libraries, logging, and pandas display — no data insights yet, just environment setup

In [1]:
```
# Import Data Manipulation Libraries
import numpy as np
import pandas as pd

# Import Data Visualization Libraries
import matplotlib.pyplot as plt
import seaborn as sns

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

# Import Logging  Files
import logging

logging.basicConfig(
    level=logging.INFO,
    filemode='w',
    filename='app.log',
    format='%(asctime)s - %(levelname)s - %(message)s')
```

In [2]:
```
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", 100)
```

Loading Data Set

In [3]:
```
# DataSet
url="https://raw.githubusercontent.com/MontyVasita18/CodeB_Internship/refs/heads/main/dataset_phishing.csv"
df=pd.read_csv(url)
df.sample(frac=1) # To make the code execution faster
```

Out[3]:

	url	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_ar
2458	http://extravasatingmetalworker.com/	36	28	0	1	0	0	0	
6281	http://apple.com.services-and-support.com/	42	34	0	3	2	0	0	
9142	http://www.medicalook.com/human_anatomy/system...	78	18	0	3	0	0	0	
4218	http://floorsdirectltd.co.uk/chase/surf4.php	44	21	0	3	0	0	0	
10325	https://wiki.creativecommons.org/wiki/public_d...	51	24	0	2	0	0	0	
...	...	...	...	...	...	...	...	...	
10893	http://www.siholding.it/gtwpages/index.jsp	42	16	0	3	0	0	0	
4504	http://nothingelsefilm.com/wp-content/themes/w...	74	19	0	1	1	0	0	
1650	https://www.havwoods.co.uk/	27	18	0	3	0	0	0	
9125	http://albex-groupe.com.ba/images/1133/interne...	70	19	0	4	1	0	0	
2069	http://wattpadsecure.000webhostapp.com/	39	31	0	2	0	0	0	

11430 rows × 89 columns

📊 Exploratory Data Analysis (EDA)

Checking shape, data types, and missing values. The dataset looks clean, with no major null-value issues. That means no need for imputation or heavy cleaning

In [4]:
```
df.shape
```

Out[4]:
```
(11430, 89)
```

In [5]:
```
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
```

In [6]:
```
# Checking Null Value in DataSet
df.isnull().sum()/len(df)*100
```

Out[6]:
```
url                         0.0
length_url                  0.0
length_hostname             0.0
ip                          0.0
nb_dots                     0.0
nb_hyphens                  0.0
nb_at                       0.0
nb_qm                       0.0
nb_and                      0.0
nb_or                       0.0
nb_eq                       0.0
nb_underscore               0.0
nb_tilde                    0.0
nb_percent                  0.0
nb_slash                    0.0
nb_star                     0.0
nb_colon                    0.0
nb_comma                    0.0
nb_semicolumn               0.0
nb_dollar                   0.0
nb_space                    0.0
nb_www                      0.0
nb_com                      0.0
nb_dslash                   0.0
http_in_path                0.0
https_token                 0.0
ratio_digits_url            0.0
ratio_digits_host           0.0
punycode                    0.0
port                        0.0
tld_in_path                 0.0
tld in subdomain            0.0
```

```
abnormal_subdomain          0.0
nb_subdomains               0.0
prefix_suffix               0.0
random_domain               0.0
shortening_service          0.0
path_extension              0.0
nb_redirection              0.0
nb_external_redirection     0.0
length_words_raw            0.0
char_repeat                 0.0
shortest_words_raw          0.0
shortest_word_host          0.0
shortest_word_path          0.0
longest_words_raw           0.0
longest_word_host           0.0
longest_word_path           0.0
avg_words_raw               0.0
avg_word_host               0.0
avg_word_path               0.0
phish_hints                 0.0
domain_in_brand             0.0
brand_in_subdomain          0.0
brand_in_path               0.0
suspecious_tld              0.0
statistical_report          0.0
nb_hyperlinks               0.0
ratio_intHyperlinks         0.0
ratio_extHyperlinks         0.0
ratio_nullHyperlinks        0.0
nb_extCSS                   0.0
ratio_intRedirection        0.0
ratio_extRedirection        0.0
ratio_intErrors             0.0
ratio_extErrors             0.0
login_form                  0.0
external_favicon            0.0
links_in_tags               0.0
submit_email                0.0
ratio_intMedia              0.0
ratio_extMedia              0.0
sfh                         0.0
iframe                      0.0
popup_window                0.0
safe_anchor                 0.0
```

```
onmouseover                 0.0
right_clic                  0.0
empty_title                 0.0
domain_in_title             0.0
domain_with_copyright       0.0
whois_registered_domain     0.0
domain_registration_length  0.0
domain_age                  0.0
web_traffic                 0.0
dns_record                  0.0
google_index                0.0
page_rank                   0.0
status                      0.0
dtype: float64
```

Summary statistics show how features are spread. You can spot skewed distributions and outliers (e.g., long URLs, high digit ratio).

In [7]:
```
df.describe()
```

Out[7]:

	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_and	nb_or
count	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	11430.000000	11430.0
mean	61.126684	21.090289	0.150569	2.480752	0.997550	0.022222	0.141207	0.162292	0.0
std	55.297318	10.777171	0.357644	1.369686	2.087087	0.155500	0.364456	0.821337	0.0
min	12.000000	4.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0
25%	33.000000	15.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.0
50%	47.000000	19.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.0
75%	71.000000	24.000000	0.000000	3.000000	1.000000	0.000000	0.000000	0.000000	0.0
max	1641.000000	214.000000	1.000000	24.000000	43.000000	4.000000	3.000000	19.000000	0.0

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 [8]:

```
# Splitting data into Numerical Data and Catagorical Data
numerical_data=df.select_dtypes(exclude='object')
numerical_data

categorical_data=df.select_dtypes(include='object')
```

In [9]:

```
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)
report=pd.DataFrame(stats)
report
```

Out[9]:

	Feature	Minimum	Maximum	Mean	Mode	25%	75%	IQR	Standard Deviation
0	length_url	12.0	1.641000e+03	61.126684	26.0	33.000000	71.000000	38.000000	5.529732e+01
1	length_hostname	4.0	2.140000e+02	21.090289	16.0	15.000000	24.000000	9.000000	1.077717e+01
2	ip	0.0	1.000000e+00	0.150569	0.0	0.000000	0.000000	0.000000	3.576436e-01
3	nb_dots	1.0	2.400000e+01	2.480752	2.0	2.000000	3.000000	1.000000	1.369686e+00
4	nb_hyphens	0.0	4.300000e+01	0.997550	0.0	0.000000	1.000000	1.000000	2.087087e+00

5	nb_at	0.0	4.000000e+00	0.022222	0.0	0.000000	0.000000	0.000000	1.554999e-01
6	nb_qm	0.0	3.000000e+00	0.141207	0.0	0.000000	0.000000	0.000000	3.644558e-01
7	nb_and	0.0	1.900000e+01	0.162292	0.0	0.000000	0.000000	0.000000	8.213374e-01
8	nb_or	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000e+00
9	nb_eq	0.0	1.900000e+01	0.293176	0.0	0.000000	0.000000	0.000000	9.983172e-01
10	nb_underscore	0.0	1.800000e+01	0.322660	0.0	0.000000	0.000000	0.000000	1.093336e+00
11	nb_tilde	0.0	1.000000e+00	0.006649	0.0	0.000000	0.000000	0.000000	8.127444e-02
12	nb_percent	0.0	9.600000e+01	0.123097	0.0	0.000000	0.000000	0.000000	1.466450e+00
13	nb_slash	2.0	3.300000e+01	4.289589	3.0	3.000000	5.000000	2.000000	1.882251e+00
14	nb_star	0.0	1.000000e+00	0.000700	0.0	0.000000	0.000000	0.000000	2.644776e-02
15	nb_colon	1.0	7.000000e+00	1.027909	1.0	1.000000	1.000000	0.000000	2.403255e-01
16	nb_comma	0.0	4.000000e+00	0.004024	0.0	0.000000	0.000000	0.000000	1.032395e-01
17	nb_semicolumn	0.0	2.000000e+01	0.062292	0.0	0.000000	0.000000	0.000000	5.981896e-01
18	nb_dollar	0.0	6.000000e+00	0.001925	0.0	0.000000	0.000000	0.000000	7.711078e-02
19	nb_space	0.0	1.800000e+01	0.034821	0.0	0.000000	0.000000	0.000000	3.755757e-01
20	nb_www	0.0	2.000000e+00	0.448469	0.0	0.000000	1.000000	1.000000	5.019124e-01
21	nb_com	0.0	6.000000e+00	0.127997	0.0	0.000000	0.000000	0.000000	3.790079e-01
22	nb_dslash	0.0	1.000000e+00	0.006562	0.0	0.000000	0.000000	0.000000	8.074153e-02
23	http_in_path	0.0	4.000000e+00	0.016710	0.0	0.000000	0.000000	0.000000	1.693581e-01
24	https_token	0.0	1.000000e+00	0.610936	1.0	0.000000	1.000000	1.000000	4.875592e-01
25	ratio_digits_url	0.0	7.238806e-01	0.053137	0.0	0.000000	0.079365	0.079365	8.936273e-02
26	ratio_digits_host	0.0	8.000000e-01	0.025024	0.0	0.000000	0.000000	0.000000	9.342200e-02
27	punycode	0.0	1.000000e+00	0.000350	0.0	0.000000	0.000000	0.000000	1.870466e-02
28	port	0.0	1.000000e+00	0.002362	0.0	0.000000	0.000000	0.000000	4.854720e-02

29	tld_in_path	0.0	1.000000e+00	0.065617	0.0	0.000000	0.000000	0.000000	2.476219e-01
30	tld_in_subdomain	0.0	1.000000e+00	0.050131	0.0	0.000000	0.000000	0.000000	2.182252e-01
31	abnormal_subdomain	0.0	1.000000e+00	0.021610	0.0	0.000000	0.000000	0.000000	1.454121e-01
32	nb_subdomains	1.0	3.000000e+00	2.231671	2.0	2.000000	3.000000	1.000000	6.370688e-01
33	prefix_suffix	0.0	1.000000e+00	0.202450	0.0	0.000000	0.000000	0.000000	4.018432e-01
34	random_domain	0.0	1.000000e+00	0.083290	0.0	0.000000	0.000000	0.000000	2.763315e-01
35	shortening_service	0.0	1.000000e+00	0.123447	0.0	0.000000	0.000000	0.000000	3.289641e-01
36	path_extension	0.0	1.000000e+00	0.000175	0.0	0.000000	0.000000	0.000000	1.322735e-02
37	nb_redirection	0.0	6.000000e+00	0.498250	0.0	0.000000	1.000000	1.000000	6.919070e-01
38	nb_external_redirection	0.0	1.000000e+00	0.003150	0.0	0.000000	0.000000	0.000000	5.603535e-02
39	length_words_raw	1.0	1.060000e+02	6.232808	2.0	2.000000	8.000000	6.000000	5.572355e+00
40	char_repeat	0.0	1.460000e+02	2.927472	3.0	1.000000	4.000000	3.000000	4.768936e+00
41	shortest_words_raw	1.0	3.100000e+01	3.127297	3.0	2.000000	3.000000	1.000000	2.211571e+00
42	shortest_word_host	1.0	3.900000e+01	5.019773	3.0	3.000000	6.000000	3.000000	3.941580e+00
43	shortest_word_path	0.0	4.000000e+01	2.398950	0.0	0.000000	3.000000	3.000000	2.997809e+00
44	longest_words_raw	2.0	8.290000e+02	15.393876	9.0	9.000000	16.000000	7.000000	2.208364e+01
45	longest_word_host	1.0	6.200000e+01	10.467979	9.0	7.000000	13.000000	6.000000	4.932015e+00
46	longest_word_path	0.0	8.290000e+02	10.561505	0.0	0.000000	11.000000	11.000000	2.307788e+01
47	avg_words_raw	2.0	1.282500e+02	7.258882	6.0	5.250000	8.000000	2.750000	4.145827e+00
48	avg_word_host	1.0	3.900000e+01	7.678075	5.0	5.250000	9.000000	3.750000	3.578435e+00
49	avg_word_path	0.0	2.500000e+02	5.092425	0.0	0.000000	6.714286	6.714286	7.147050e+00
50	phish_hints	0.0	1.000000e+01	0.327734	0.0	0.000000	0.000000	0.000000	8.426004e-01
51	domain_in_brand	0.0	1.000000e+00	0.104199	0.0	0.000000	0.000000	0.000000	3.055325e-01
52	brand_in_subdomain	0.0	1.000000e+00	0.004112	0.0	0.000000	0.000000	0.000000	6.399559e-02

53	brand_in_path	0.0	1.000000e+00	0.004899	0.0	0.000000	0.000000	0.000000	6.982700e-02
54	suspecious_tld	0.0	1.000000e+00	0.017935	0.0	0.000000	0.000000	0.000000	1.327220e-01
55	statistical_report	0.0	2.000000e+00	0.059755	0.0	0.000000	0.000000	0.000000	3.312662e-01
56	nb_hyperlinks	0.0	4.659000e+03	87.189764	0.0	9.000000	101.000000	92.000000	1.667583e+02
57	ratio_intHyperlinks	0.0	1.000000e+00	0.602457	0.0	0.224991	0.944767	0.719776	3.764745e-01
58	ratio_extHyperlinks	0.0	1.000000e+00	0.276720	0.0	0.000000	0.474840	0.474840	3.199583e-01
59	ratio_nullHyperlinks	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000e+00
60	nb_extCSS	0.0	1.240000e+02	0.784864	0.0	0.000000	1.000000	1.000000	2.758802e+00
61	ratio_intRedirection	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000e+00
62	ratio_extRedirection	0.0	2.000000e+00	0.158926	0.0	0.000000	0.230769	0.230769	2.664370e-01
63	ratio_intErrors	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000e+00
64	ratio_extErrors	0.0	1.000000e+00	0.062469	0.0	0.000000	0.034483	0.034483	1.562087e-01
65	login_form	0.0	1.000000e+00	0.063605	0.0	0.000000	0.000000	0.000000	2.440578e-01
66	external_favicon	0.0	1.000000e+00	0.442170	0.0	0.000000	1.000000	1.000000	4.966661e-01
67	links_in_tags	0.0	1.000000e+02	51.978211	0.0	0.000000	98.061004	98.061004	4.152314e+01
68	submit_email	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000e+00
69	ratio_intMedia	0.0	1.000000e+02	42.870444	0.0	0.000000	100.000000	100.000000	4.624990e+01
70	ratio_extMedia	0.0	1.000000e+02	23.236293	0.0	0.000000	33.333333	33.333333	3.838658e+01
71	sfh	0.0	0.000000e+00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000e+00
72	iframe	0.0	1.000000e+00	0.001312	0.0	0.000000	0.000000	0.000000	3.620398e-02
73	popup_window	0.0	1.000000e+00	0.006037	0.0	0.000000	0.000000	0.000000	7.746501e-02
74	safe_anchor	0.0	1.000000e+02	37.063922	0.0	0.000000	75.000000	75.000000	3.907339e+01
75	onmouseover	0.0	1.000000e+00	0.001137	0.0	0.000000	0.000000	0.000000	3.370703e-02
76	right_clic	0.0	1.000000e+00	0.001400	0.0	0.000000	0.000000	0.000000	3.738968e-02

77	empty_title	0.0	1.000000e+00	0.124759	0.0	0.000000	0.000000	0.000000	3.304604e-01
78	domain_in_title	0.0	1.000000e+00	0.775853	1.0	1.000000	1.000000	0.000000	4.170376e-01
79	domain_with_copyright	0.0	1.000000e+00	0.439545	0.0	0.000000	1.000000	1.000000	4.963535e-01
80	whois_registered_domain	0.0	1.000000e+00	0.072878	0.0	0.000000	0.000000	0.000000	2.599482e-01
81	domain_registration_length	-1.0	2.982900e+04	492.532196	0.0	84.000000	449.000000	365.000000	8.147694e+02
82	domain_age	-12.0	1.287400e+04	4062.543745	-1.0	972.250000	7026.750000	6054.500000	3.107785e+03
83	web_traffic	0.0	1.076799e+07	856756.643307	0.0	0.000000	373845.500000	373845.500000	1.995606e+06
84	dns_record	0.0	1.000000e+00	0.020122	0.0	0.000000	0.000000	0.000000	1.404254e-01
85	google_index	0.0	1.000000e+00	0.533946	1.0	0.000000	1.000000	1.000000	4.988682e-01
86	page_rank	0.0	1.000000e+01	3.185739	0.0	1.000000	5.000000	4.000000	2.536955e+00

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 [10]:
```
# Frequency Distribution
for col in df.columns:
    if df[col].dtype=='object':
        print(f"Frequency Distribution Of {col}\n")
        print(df[col].value_counts)
```

```
Frequency Distribution Of url

<bound method IndexOpsMixin.value_counts of 0                    http://www.crestonwood.com/router.php
1        http://shadetreetechnology.com/V4/validation/a...
2        https://support-appleld.com.secureupdate.duila...
3                                     http://rgipt.ac.in
```

```
4        http://www.iracing.com/tracks/gateway-motorspo...
                                  ...
11425        http://www.fontspace.com/category/blackletter
11426    http://www.budgetbots.com/server.php/Server%20...
11427    https://www.facebook.com/Interactive-Televisio...
11428               http://www.mypublicdomainpictures.com/
11429    http://174.139.46.123/ap/signin?openid.pape.ma...
Name: url, Length: 11430, dtype: object>
Frequency Distribution Of status

<bound method IndexOpsMixin.value_counts of 0        legitimate
1          phishing
2          phishing
3        legitimate
4        legitimate
            ...
11425    legitimate
11426      phishing
11427    legitimate
11428    legitimate
11429      phishing
Name: status, Length: 11430, dtype: object>
```

In [11]:
```
df['status'].value_counts()
```

Out[11]:
```
status
legitimate    5715
phishing      5715
Name: count, dtype: int64
```

The target column status is well-balanced, which is ideal for binary classification models and ensures fair learning across both classes.

In [12]:
```
df['status'].mode()
```

Out[12]:
```
0    legitimate
1      phishing
Name: status, dtype: object
```

Label encoding turns 'legitimate' and 'phishing' into 0 and 1 — readying the target for machine learning models.

In [13]:
```
# Encoding Target column
```

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

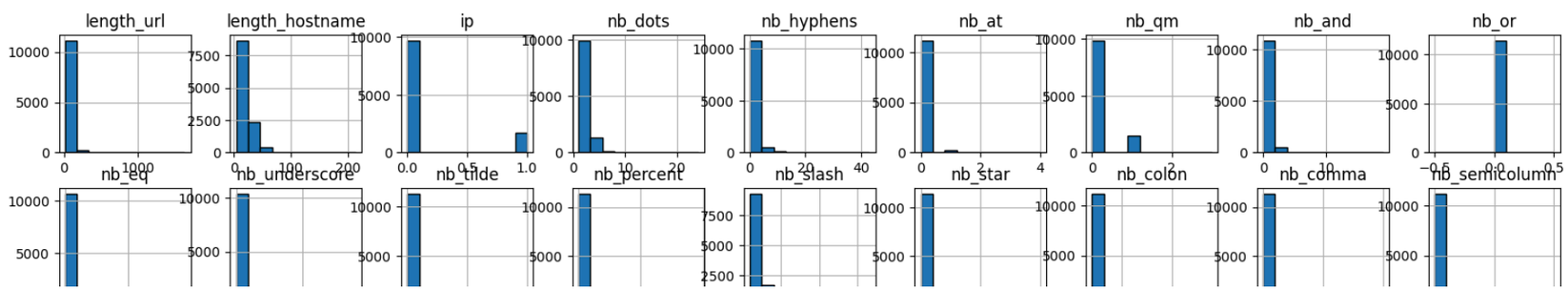
Out[13]:
```
0        0
1        1
2        1
3        0
4        0
        ..
11425    0
11426    1
11427    0
11428    0
11429    1
Name: status, Length: 11430, dtype: int64
```

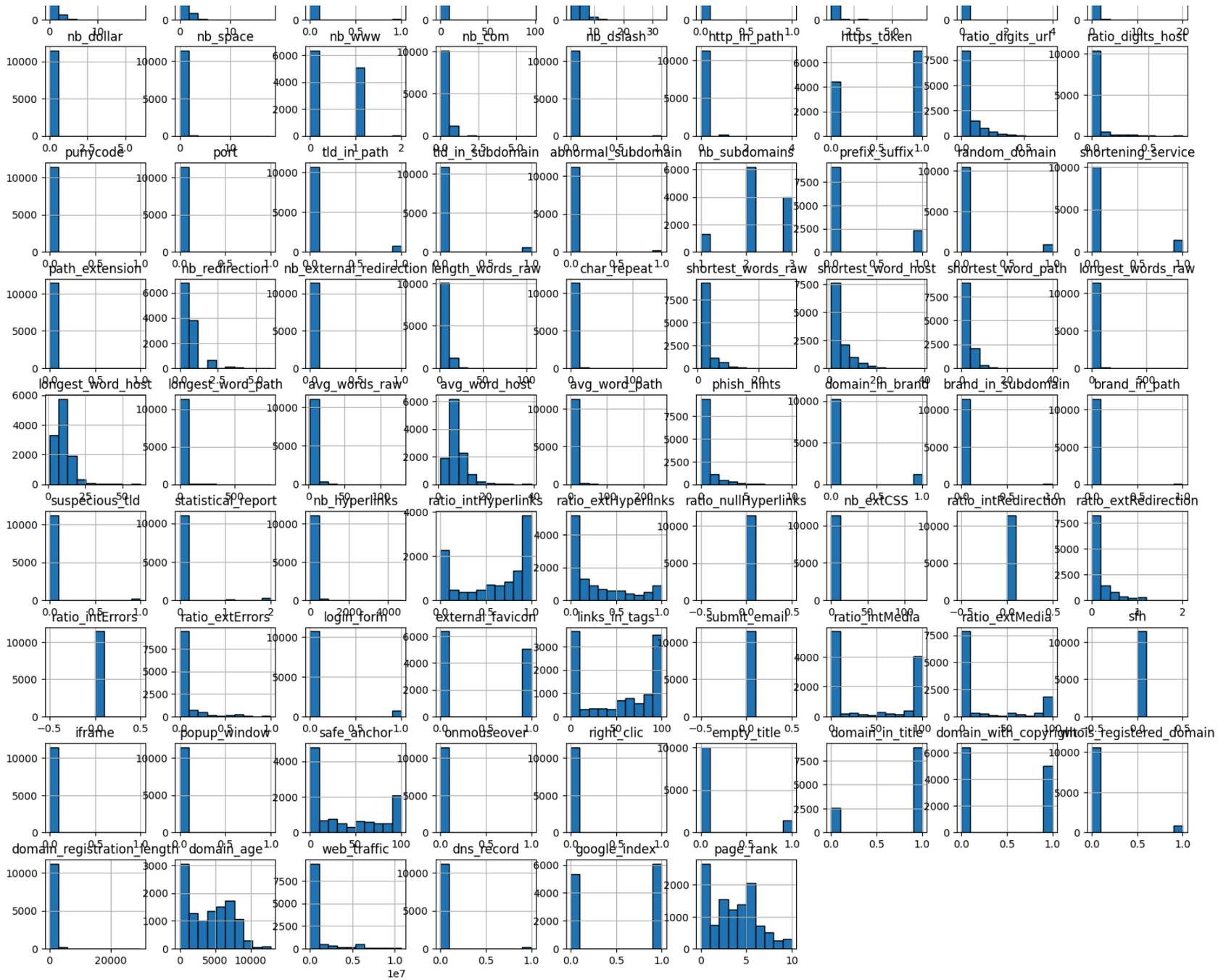
The target variable status was originally categorical, labeled as “phishing” and “legitimate.” It was converted into a binary format (1 and 0) for model compatibility.

Histogram

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

In [14]:
```
# Plotting Histogram
numerical_data.hist(figsize=(20,20),bins=10,edgecolor='black')
plt.title('Histogram example',y=1.02)
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```

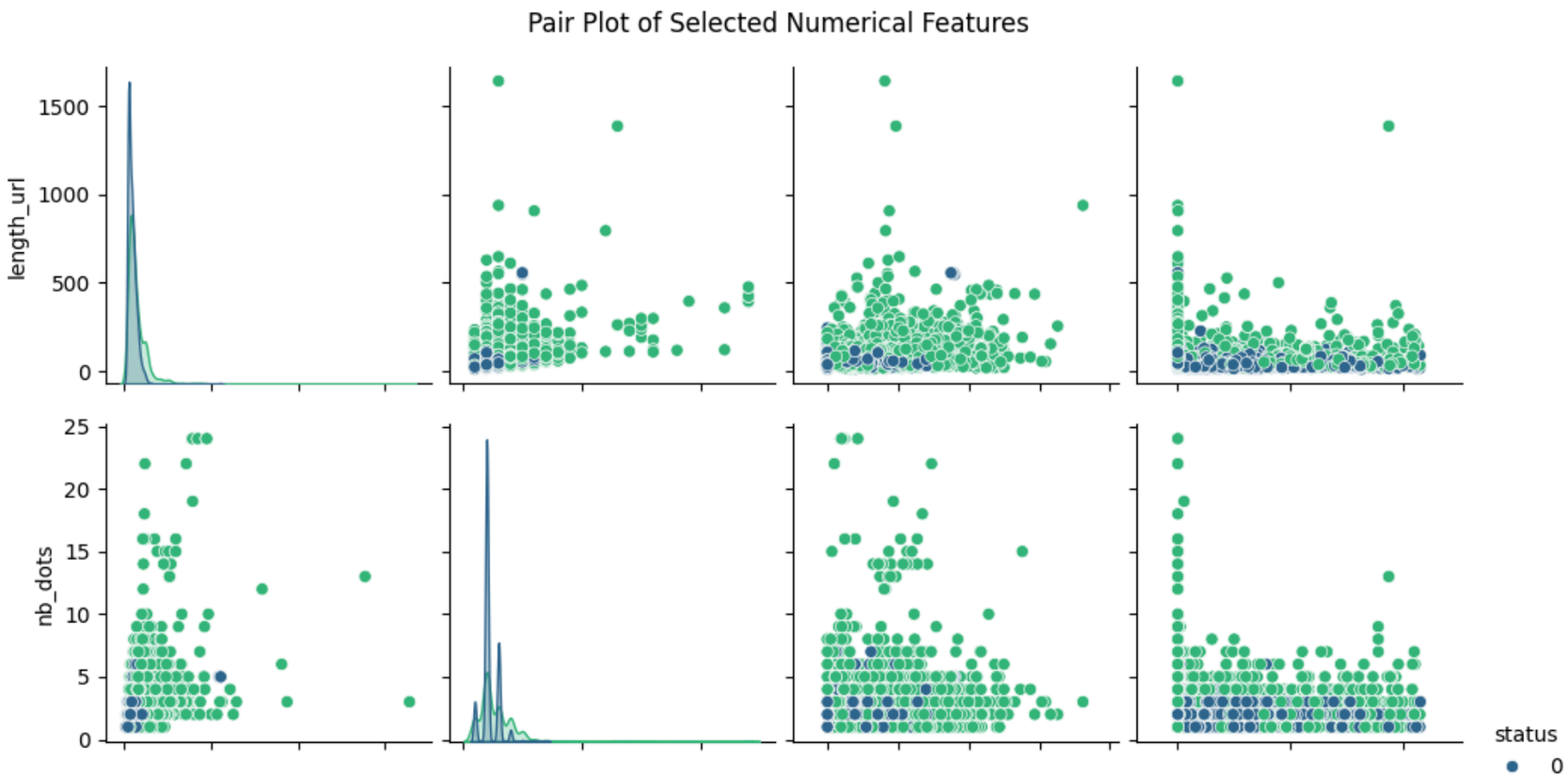


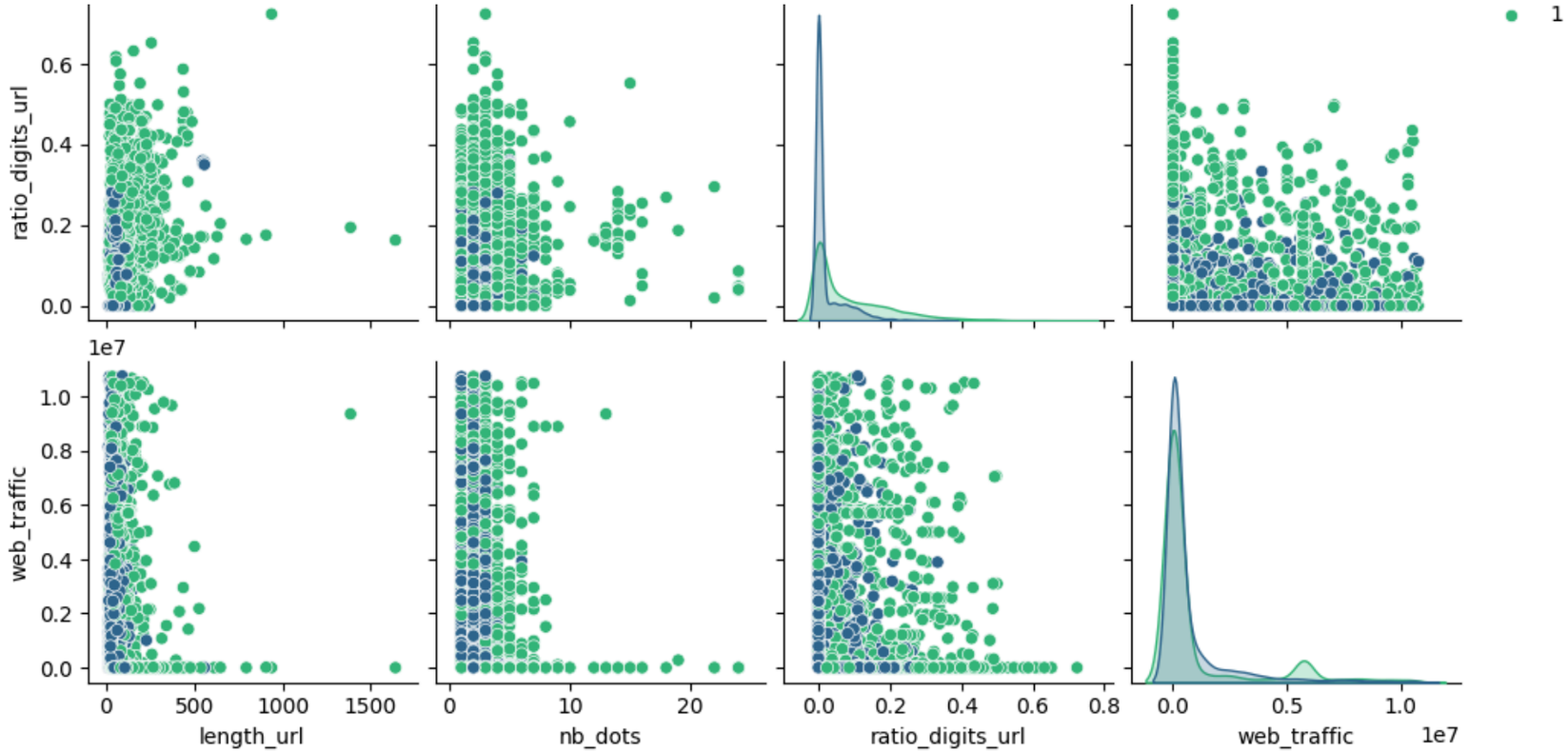

nb_dollar
nb_space
nb_www
nb_com
nb_dslash
http_in_path
https_token
ratio_digits_url
ratio_digits_host
punycode
port
tld_in_path
tld_in_subdomain
abnormal_subdomain
nb_subdomains
prefix_suffix
random_domain
shortening_service
path_extension
nb_redirection
nb_external_redirection
length_words_raw
char_repeat
shortest_words_raw
shortest_word_host
shortest_word_path
longest_words_raw
longest_word_host
longest_word_path
avg_words_raw
avg_word_host
avg_word_path
phish_hints
domain_in_brand
brand_in_subdomain
brand_in_path
suspecious_tld
statistical_report
nb_hyperlinks
ratio_intHyperlinks
ratio_extHyperlinks
ratio_nullHyperlinks
nb_extCSS
ratio_intRedirection
ratio_extRedirection
ratio_intErrors
ratio_extErrors
login_form
external_favicon
links_in_tags
submit_email
ratio_intMedia
ratio_extMedia
sfh
iframe
popup_window
safe_anchor
onmouseover
right_clic
empty_title
domain_in_title
domain_with_copyright
whois_registered_domain
domain_registration_length
domain_age
web_traffic
dns_record
google_index
page_rank
1e7

Many features are right-skewed, indicating potential preprocessing needs (e.g., log transformation). Distribution plots also highlighted high concentration of values in specific ranges for features like ratio_digits_url.

Pair Plot

In [15]:
```
selected_features = ['length_url', 'nb_dots', 'ratio_digits_url', 'web_traffic', 'status']
# Plot pair plot
sns.pairplot(df[selected_features], hue='status', palette='viridis')
# Optional: Add title
plt.suptitle("Pair Plot of Selected Numerical Features", y=1.02)
plt.show()
```





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.

✅ Insights & Recommendations

Key Findings:

Several numerical features display non-normal distributions and contain outliers, which could affect model performance if not addressed.

Features like ratio_digits_url and web_traffic show clear separation between classes and can act as strong indicators for phishing detection.

The target column status is well-balanced, which is ideal for binary classification models and ensures fair learning across both classes

The target column 'status' is well-balanced, which is ideal for binary classification models and ensures fair learning across both classes.

Recommended Actions:

Normalize or transform skewed numerical features (e.g., using log or power transforms) to reduce the effect of extreme values.

Scale features using standardization (e.g., MinMaxScaler or StandardScaler) to ensure uniform treatment by algorithms.

Use feature selection techniques (e.g., correlation thresholding, mutual information, or tree-based feature importance) to focus on the most predictive variables.

Check for multicollinearity using correlation matrices or VIF to avoid redundant features

In [ ]:

Checking duplicates

In [19]:
```
duplicates=df.duplicated()
```

In [21]:
```
duplicates.value_counts()
```

Out[21]:
```
False    11430
Name: count, dtype: int64
```

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 [24]:
```
# label Encoding Url Column

from sklearn.preprocessing import LabelEncoder
LE=LabelEncoder()
df['url']=LE.fit_transform(df['url'])
```

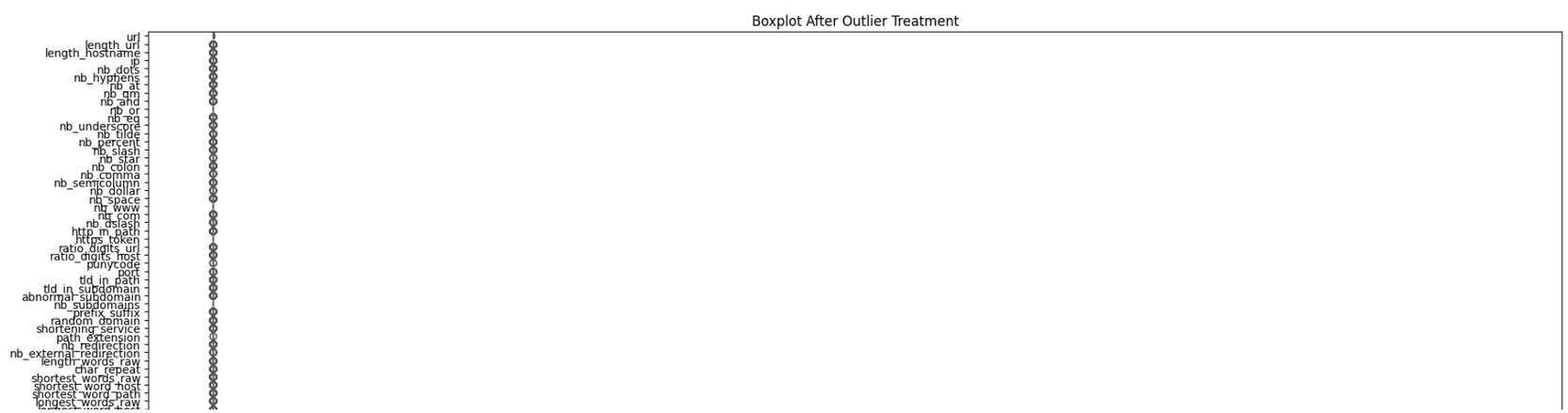
```
df['url'].value_counts()
```

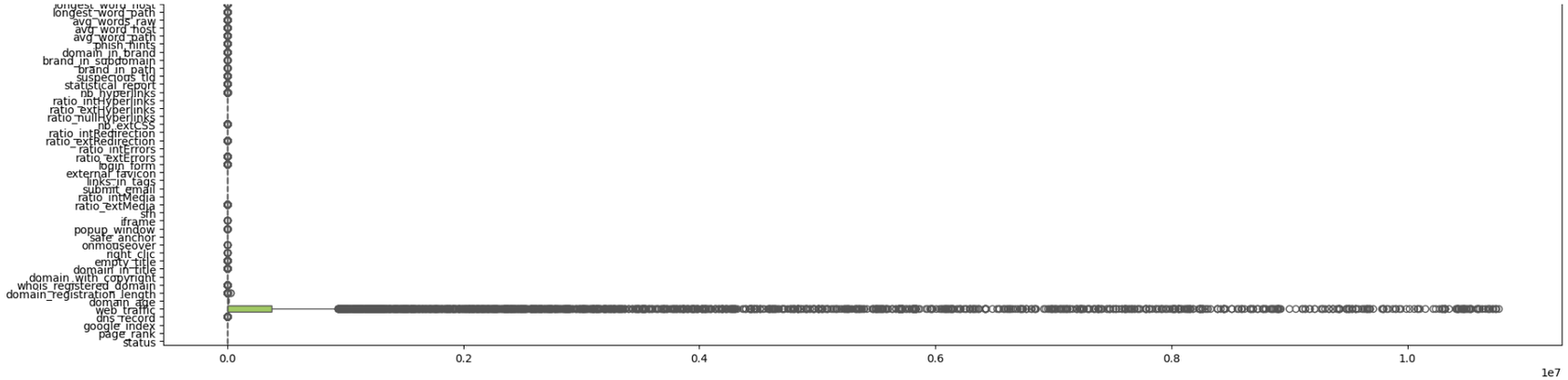
Out[24]:
```
url
1065     2
4501     1
10779    1
1315     1
9201     1
        ..
6539     1
819      1
9629     1
5956     1
62       1
Name: count, Length: 11429, dtype: int64
```

In [29]:
```
# Checking Outliers Using Boxplot
# Set figure size
plt.figure(figsize=(20, 10))

# 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()
```





In [34]:
```
# Spliting Data into Independent And target Column
X=df.drop(columns='status')
y=df['status']
```

In [35]:
```
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 [36]:
```
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 [ ]:
```
from sklearn.preprocessing import MinMaxScaler,StandardScaler,RobustScaler
scaler=RobustScaler()
X_train=scaler.fit_transform(X_train)
X_test=scaler.transform(X_test)
```

In [39]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

```
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()
```