```
# 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')

pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', 100)
```

Loading Dataset

```
# Loading the dataset

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

df = pd.read_csv(url)

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

```
                                                      url
length_url  \
7039                      http://www.lastingredient.com/          30

9187   http://mail.midyatmimaritas.com/wp-includes/im...          70

1935   https://zoomic.io/wp-includes/neworder/bizmail...         145

4915                    http://www.starwalkerstudios.com/          33

5670   http://www.auth-chaseuserservice.ssmailer.com/...          72
```

```
...                                                   ...          ...

2782      http://shiflett.org/articles/session-hijacking           46

11137                       http://www.tattooartists.ru/           28

1535   http://yovcxm.com/chase/Chase/ef195f766730a094...           62

7845        http://www.yourdictionary.com/standard-error           44

9078   https://bugzilla.redhat.com/show_bug.cgi?id=90...           50


       length_hostname  ip  nb_dots  nb_hyphens  nb_at  nb_qm  nb_and
nb_or  \
7039                22   0        2           0      0      0       0
0
9187                24   0        2           1      0      0       0
0
1935                 9   0        2           2      0      1       6
0
4915                25   0        2           0      0      0       0
0
5670                38   0        4           1      0      0       0
0
...                ...  ..      ...         ...    ...    ...     ...
...
2782                12   0        1           1      0      0       0
0
11137               20   0        2           0      0      0       0
0
1535                10   1        1           0      0      0       0
0
7845                22   0        2           1      0      0       0
0
9078                19   0        3           0      0      1       0
0

       nb_eq  nb_underscore  nb_tilde  nb_percent  nb_slash
nb_star  \
7039       0              0         0           0         3        0

9187       0              0         0           0         7        0

1935       7              1         0           0         5        0

4915       0              0         0           0         3        0

5670       0              0         0           0         6        0
```

```
...      ...             ...       ...          ...       ...      ...
2782       0               0         0            0         4        0
11137      0               0         0            0         3        0
1535       0               0         0            0         5        0
7845       0               0         0            0         3        0
9078       1               1         0            0         3        0


       nb_colon  nb_comma  nb_semicolumn  nb_dollar  nb_space  nb_www
nb_com  \
7039          1         0              0          0         0       1
0
9187          1         0              0          0         0       0
0
1935          1         0              6          0         0       0
0
4915          1         0              0          0         0       1
0
5670          1         0              0          0         0       1
0
...         ...       ...            ...        ...       ...     ...
...
2782          1         0              0          0         0       0
0
11137         1         0              0          0         0       1
0
1535          1         0              0          0         0       0
0
7845          1         0              0          0         0       1
0
9078          1         0              0          0         0       0
0

       nb_dslash  http_in_path  https_token  ratio_digits_url  \
7039           0             0            1          0.000000
9187           0             0            1          0.057143
1935           0             0            0          0.096552
4915           0             0            1          0.000000
5670           0             0            1          0.000000
...          ...           ...          ...               ...
2782           0             0            1          0.000000
11137          0             0            1          0.000000
1535           0             0            1          0.370968
7845           0             0            1          0.000000
9078           0             0            0          0.120000
```

```
       ratio_digits_host  punycode  port  tld_in_path
tld_in_subdomain  \
7039                 0.0         0     0            0
0
9187                 0.0         0     0            0
0
1935                 0.0         0     0            0
0
4915                 0.0         0     0            0
0
5670                 0.0         0     0            0
0
...                  ...       ...   ...          ...                ..
.
2782                 0.0         0     0            0
0
11137                0.0         0     0            0
0
1535                 0.0         0     0            0
0
7845                 0.0         0     0            0
0
9078                 0.0         0     0            0
0

       abnormal_subdomain  nb_subdomains  prefix_suffix  random_domain
\
7039                    0              2              0              0

9187                    0              2              1              0

1935                    0              2              1              0

4915                    0              2              0              0

5670                    0              3              1              0

...                   ...            ...            ...            ...

2782                    0              1              0              0

11137                   0              2              0              0

1535                    0              1              0              1

7845                    0              2              0              0

9078                    0              3              0              0
```

```
       shortening_service  path_extension  nb_redirection  \
7039                    1               0               1
9187                    0               0               1
1935                    0               0               0
4915                    0               0               0
5670                    0               0               1
...                   ...             ...             ...
2782                    0               0               1
11137                   0               0               0
1535                    0               0               1
7845                    0               0               1
9078                    1               0               0

       nb_external_redirection  length_words_raw  char_repeat  \
7039                         0                 2            3
9187                         0                 8            0
1935                         0                21            3
4915                         0                 2            3
5670                         0                 9            4
...                        ...               ...          ...
2782                         0                 4            2
11137                        0                 2            5
1535                         0                 4            1
7845                         0                 4            4
9078                         0                 7            1

       shortest_words_raw  shortest_word_host  shortest_word_path  \
7039                    3                   3                   0
9187                    2                   4                   2
1935                    1                   6                   1
4915                    3                   3                   0
5670                    3                   3                   3
...                   ...                 ...                 ...
2782                    7                   8                   7
11137                   3                   3                   0
1535                    5                   6                   5
7845                    3                   3                   5
9078                    2                   6                   2

       longest_words_raw  longest_word_host  longest_word_path
avg_words_raw  \
7039                  14                 14                  0
8.500000
9187                  15                 15                  8
6.500000
1935                  13                  6                 13
5.380952
4915                  17                 17                  0
10.000000
5670                  16                 16                  6
```

```
5.888889
...                    ...                  ...                  ...
...
2782                    9                    8                    9
8.000000
11137                  13                   13                    0
8.000000
1535                   32                    6                   32
12.000000
7845                   14                   14                    8
7.500000
9078                    8                    8                    6
4.571429

       avg_word_host  avg_word_path  phish_hints  domain_in_brand  \
7039            8.50           0.00            0                 0
9187            9.50           5.50            3                 0
1935            6.00           5.35            2                 0
4915           10.00           0.00            0                 0
5670            7.75           4.40            0                 0
...              ...            ...          ...               ...
2782            8.00           8.00            0                 0
11137           8.00           0.00            0                 0
1535            6.00          14.00            0                 0
7845            8.50           6.50            0                 0
9078            7.00           3.60            0                 0

       brand_in_subdomain  brand_in_path  suspecious_tld
statistical_report  \
7039                    0              0               0
0
9187                    0              0               0
0
1935                    0              0               0
0
4915                    0              0               0
0
5670                    0              0               0
0
...                   ...            ...             ...
...
2782                    0              0               0
0
11137                   0              0               0
0
1535                    0              0               0
0
7845                    0              0               0
0
```

```
9078                    0              0                0
0

       nb_hyperlinks  ratio_intHyperlinks  ratio_extHyperlinks  \
7039              41             0.585366             0.414634
9187              38             0.421053             0.578947
1935               8             1.000000             0.000000
4915             104             0.903846             0.096154
5670             103             0.932039             0.067961
...              ...                  ...                  ...
2782              48             0.958333             0.041667
11137             88             0.920455             0.079545
1535              48             1.000000             0.000000
7845             195             0.994872             0.005128
9078             109             1.000000             0.000000

       ratio_nullHyperlinks  nb_extCSS  ratio_intRedirection  \
7039                      0          1                     0
9187                      0          2                     0
1935                      0          0                     0
4915                      0          0                     0
5670                      0          1                     0
...                     ...        ...                   ...
2782                      0          0                     0
11137                     0          0                     0
1535                      0          0                     0
7845                      0          0                     0
9078                      0          0                     0

       ratio_extRedirection  ratio_intErrors  ratio_extErrors
login_form  \
7039               0.000000                0         0.117647
0
9187               0.227273                0         0.181818
0
1935               0.000000                0         0.000000
0
4915               0.200000                0         0.000000
0
5670               0.142857                0         0.000000
0
...                     ...              ...              ...
...
2782               0.000000                0         0.000000
0
11137              0.285714                0         0.000000
0
1535               0.000000                0         0.000000
0
7845               0.000000                0         0.000000
```

```
0
9078                0.000000                0         0.000000
0

       external_favicon  links_in_tags  submit_email
ratio_intMedia  \
7039                  1      61.538462             0      100.000000

9187                  1      42.857143             0        0.000000

1935                  0     100.000000             0      100.000000

4915                  1      72.727273             0       42.857143

5670                  1      81.818182             0      100.000000

...                 ...            ...           ...             ...

2782                  1      92.857143             0      100.000000

11137                 0      50.000000             0       92.592593

1535                  0     100.000000             0      100.000000

7845                  0     100.000000             0      100.000000

9078                  0     100.000000             0      100.000000


       ratio_extMedia  sfh  iframe  popup_window  safe_anchor
onmouseover  \
7039         0.000000    0       0             0    11.111111
0
9187       100.000000    0       0             0    11.111111
0
1935         0.000000    0       0             0     0.000000
0
4915        57.142857    0       0             0   100.000000
0
5670         0.000000    0       0             0     0.000000
0
...               ...  ...     ...           ...          ...
...
2782         0.000000    0       0             0   100.000000
0
11137        7.407407    0       0             0    66.666667
0
1535         0.000000    0       0             0   100.000000
0
7845         0.000000    0       0             0    93.750000
```

```
0
9078         0.000000    0       0             0   100.000000
0

       right_clic  empty_title  domain_in_title  domain_with_copyright
\
7039            0            0                1                      1

9187            0            0                1                      1

1935            0            0                1                      0

4915            0            0                1                      1

5670            0            0                1                      1

...           ...          ...              ...                    ...

2782            0            0                0                      0

11137           0            0                1                      0

1535            0            0                1                      1

7845            0            0                1                      1

9078            0            0                1                      0


       whois_registered_domain  domain_registration_length  domain_age
\
7039                         0                         358        3295

9187                         0                         318         778

1935                         0                          58          -1

4915                         0                         135        2421

5670                         0                         351        5858

...                        ...                         ...         ...

2782                         0                         649        7386

11137                        0                         115        4633

1535                         0                         246          -1

7845                         0                         449        7586
```

```
9078                          0                           306        9554


       web_traffic  dns_record  google_index  page_rank      status
7039       1759878           0             0          4  legitimate
9187             0           0             1          0    phishing
1935      10687767           0             1          0    phishing
4915             0           0             0          4  legitimate
5670             0           0             1          0    phishing
...            ...         ...           ...        ...         ...
2782       1213116           0             0          5  legitimate
11137      3147040           0             0          2  legitimate
1535             0           0             1          0    phishing
7845           982           0             0          5  legitimate
9078          4350           0             0          7  legitimate

[11430 rows x 89 columns]
```

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

```
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

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

```
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

```
df.describe()
```

```
         length_url  length_hostname            ip       nb_dots  \
count  11430.000000     11430.000000  11430.000000  11430.000000
mean      61.126684        21.090289      0.150569      2.480752
std       55.297318        10.777171      0.357644      1.369686
min       12.000000         4.000000      0.000000      1.000000
25%       33.000000        15.000000      0.000000      2.000000
50%       47.000000        19.000000      0.000000      2.000000
75%       71.000000        24.000000      0.000000      3.000000
max     1641.000000       214.000000      1.000000     24.000000

         nb_hyphens         nb_at         nb_qm        nb_and    nb_or
\
count  11430.000000  11430.000000  11430.000000  11430.000000  11430.0

mean       0.997550      0.022222      0.141207      0.162292      0.0

std        2.087087      0.155500      0.364456      0.821337      0.0

min        0.000000      0.000000      0.000000      0.000000      0.0
```

```
25%         0.000000      0.000000      0.000000      0.000000       0.0

50%         0.000000      0.000000      0.000000      0.000000       0.0

75%         1.000000      0.000000      0.000000      0.000000       0.0

max        43.000000      4.000000      3.000000     19.000000       0.0


               nb_eq  nb_underscore      nb_tilde    nb_percent
nb_slash  \
count  11430.000000   11430.000000  11430.000000  11430.000000
11430.000000
mean       0.293176       0.322660      0.006649      0.123097
4.289589
std        0.998317       1.093336      0.081274      1.466450
1.882251
min        0.000000       0.000000      0.000000      0.000000
2.000000
25%        0.000000       0.000000      0.000000      0.000000
3.000000
50%        0.000000       0.000000      0.000000      0.000000
4.000000
75%        0.000000       0.000000      0.000000      0.000000
5.000000
max       19.000000      18.000000      1.000000     96.000000
33.000000

             nb_star      nb_colon      nb_comma  nb_semicolumn
nb_dollar  \
count  11430.000000  11430.000000  11430.000000   11430.000000
11430.000000
mean       0.000700      1.027909      0.004024       0.062292
0.001925
std        0.026448      0.240325      0.103240       0.598190
0.077111
min        0.000000      1.000000      0.000000       0.000000
0.000000
25%        0.000000      1.000000      0.000000       0.000000
0.000000
50%        0.000000      1.000000      0.000000       0.000000
0.000000
75%        0.000000      1.000000      0.000000       0.000000
0.000000
max        1.000000      7.000000      4.000000      20.000000
6.000000

           nb_space        nb_www        nb_com     nb_dslash
http_in_path  \
count  11430.000000  11430.000000  11430.000000  11430.000000
```

```
11430.000000
mean       0.034821      0.448469      0.127997      0.006562
0.016710
std        0.375576      0.501912      0.379008      0.080742
0.169358
min        0.000000      0.000000      0.000000      0.000000
0.000000
25%        0.000000      0.000000      0.000000      0.000000
0.000000
50%        0.000000      0.000000      0.000000      0.000000
0.000000
75%        0.000000      1.000000      0.000000      0.000000
0.000000
max       18.000000      2.000000      6.000000      1.000000
4.000000

        https_token  ratio_digits_url  ratio_digits_host      punycode
\
count  11430.000000      11430.000000       11430.000000  11430.000000

mean       0.610936          0.053137           0.025024      0.000350

std        0.487559          0.089363           0.093422      0.018705

min        0.000000          0.000000           0.000000      0.000000

25%        0.000000          0.000000           0.000000      0.000000

50%        1.000000          0.000000           0.000000      0.000000

75%        1.000000          0.079365           0.000000      0.000000

max        1.000000          0.723881           0.800000      1.000000


                port   tld_in_path  tld_in_subdomain
abnormal_subdomain  \
count  11430.000000  11430.000000      11430.000000
11430.000000
mean       0.002362      0.065617          0.050131
0.021610
std        0.048547      0.247622          0.218225
0.145412
min        0.000000      0.000000          0.000000
0.000000
25%        0.000000      0.000000          0.000000
0.000000
50%        0.000000      0.000000          0.000000
0.000000
75%        0.000000      0.000000          0.000000
```

```
0.000000
max        1.000000      1.000000           1.000000
1.000000

       nb_subdomains  prefix_suffix  random_domain  shortening_service
\
count   11430.000000   11430.000000   11430.000000        11430.000000

mean        2.231671       0.202450       0.083290            0.123447

std         0.637069       0.401843       0.276332            0.328964

min         1.000000       0.000000       0.000000            0.000000

25%         2.000000       0.000000       0.000000            0.000000

50%         2.000000       0.000000       0.000000            0.000000

75%         3.000000       0.000000       0.000000            0.000000

max         3.000000       1.000000       1.000000            1.000000


       path_extension  nb_redirection  nb_external_redirection  \
count    11430.000000    11430.000000             11430.000000
mean         0.000175        0.498250                 0.003150
std          0.013227        0.691907                 0.056035
min          0.000000        0.000000                 0.000000
25%          0.000000        0.000000                 0.000000
50%          0.000000        0.000000                 0.000000
75%          0.000000        1.000000                 0.000000
max          1.000000        6.000000                 1.000000

       length_words_raw   char_repeat  shortest_words_raw
shortest_word_host  \
count      11430.000000  11430.000000        11430.000000
11430.000000
mean           6.232808      2.927472            3.127297
5.019773
std            5.572355      4.768936            2.211571
3.941580
min            1.000000      0.000000            1.000000
1.000000
25%            2.000000      1.000000            2.000000
3.000000
50%            5.000000      3.000000            3.000000
3.000000
75%            8.000000      4.000000            3.000000
6.000000
max          106.000000    146.000000           31.000000
```

```
39.000000

       shortest_word_path  longest_words_raw  longest_word_host  \
count        11430.000000       11430.000000       11430.000000
mean             2.398950          15.393876          10.467979
std              2.997809          22.083644           4.932015
min              0.000000           2.000000           1.000000
25%              0.000000           9.000000           7.000000
50%              2.000000          11.000000          10.000000
75%              3.000000          16.000000          13.000000
max             40.000000         829.000000          62.000000

       longest_word_path  avg_words_raw  avg_word_host  avg_word_path
\
count       11430.000000   11430.000000   11430.000000   11430.000000

mean           10.561505       7.258882       7.678075       5.092425

std            23.077883       4.145827       3.578435       7.147050

min             0.000000       2.000000       1.000000       0.000000

25%             0.000000       5.250000       5.250000       0.000000

50%             7.000000       6.500000       7.000000       4.857143

75%            11.000000       8.000000       9.000000       6.714286

max           829.000000     128.250000      39.000000     250.000000


        phish_hints  domain_in_brand  brand_in_subdomain
brand_in_path  \
count  11430.000000     11430.000000        11430.000000
11430.000000
mean       0.327734         0.104199            0.004112
0.004899
std        0.842600         0.305533            0.063996
0.069827
min        0.000000         0.000000            0.000000
0.000000
25%        0.000000         0.000000            0.000000
0.000000
50%        0.000000         0.000000            0.000000
0.000000
75%        0.000000         0.000000            0.000000
0.000000
max       10.000000         1.000000            1.000000
1.000000
```

```
       suspecious_tld  statistical_report  nb_hyperlinks
ratio_intHyperlinks  \
count    11430.000000        11430.000000   11430.000000
11430.000000
mean         0.017935            0.059755      87.189764
0.602457
std          0.132722            0.331266     166.758254
0.376474
min          0.000000            0.000000       0.000000
0.000000
25%          0.000000            0.000000       9.000000
0.224991
50%          0.000000            0.000000      34.000000
0.743442
75%          0.000000            0.000000     101.000000
0.944767
max          1.000000            2.000000    4659.000000
1.000000

       ratio_extHyperlinks  ratio_nullHyperlinks     nb_extCSS  \
count         11430.000000               11430.0  11430.000000
mean              0.276720                   0.0      0.784864
std               0.319958                   0.0      2.758802
min               0.000000                   0.0      0.000000
25%               0.000000                   0.0      0.000000
50%               0.131148                   0.0      0.000000
75%               0.474840                   0.0      1.000000
max               1.000000                   0.0    124.000000

       ratio_intRedirection  ratio_extRedirection  ratio_intErrors  \
count               11430.0          11430.000000          11430.0
mean                    0.0              0.158926              0.0
std                     0.0              0.266437              0.0
min                     0.0              0.000000              0.0
25%                     0.0              0.000000              0.0
50%                     0.0              0.000000              0.0
75%                     0.0              0.230769              0.0
max                     0.0              2.000000              0.0

       ratio_extErrors    login_form  external_favicon  links_in_tags
\
count     11430.000000  11430.000000      11430.000000   11430.000000

mean          0.062469      0.063605          0.442170      51.978211

std           0.156209      0.244058          0.496666      41.523144

min           0.000000      0.000000          0.000000       0.000000

25%           0.000000      0.000000          0.000000       0.000000
```

```
50%           0.000000      0.000000          0.000000      60.000000

75%           0.034483      0.000000          1.000000      98.061004

max           1.000000      1.000000          1.000000     100.000000


       submit_email  ratio_intMedia  ratio_extMedia      sfh
iframe  \
count       11430.0    11430.000000    11430.000000  11430.0
11430.000000
mean            0.0       42.870444       23.236293      0.0
0.001312
std             0.0       46.249897       38.386577      0.0
0.036204
min             0.0        0.000000        0.000000      0.0
0.000000
25%             0.0        0.000000        0.000000      0.0
0.000000
50%             0.0       11.111111        0.000000      0.0
0.000000
75%             0.0      100.000000       33.333333      0.0
0.000000
max             0.0      100.000000      100.000000      0.0
1.000000

       popup_window   safe_anchor   onmouseover   right_clic
empty_title  \
count  11430.000000  11430.000000  11430.000000  11430.00000
11430.000000
mean       0.006037     37.063922      0.001137      0.00140
0.124759
std        0.077465     39.073385      0.033707      0.03739
0.330460
min        0.000000      0.000000      0.000000      0.00000
0.000000
25%        0.000000      0.000000      0.000000      0.00000
0.000000
50%        0.000000     23.294574      0.000000      0.00000
0.000000
75%        0.000000     75.000000      0.000000      0.00000
0.000000
max        1.000000    100.000000      1.000000      1.00000
1.000000

       domain_in_title  domain_with_copyright  whois_registered_domain
\
count     11430.000000           11430.000000             11430.000000
```

```
mean          0.775853                0.439545                0.072878

std           0.417038                0.496353                0.259948

min           0.000000                0.000000                0.000000

25%           1.000000                0.000000                0.000000

50%           1.000000                0.000000                0.000000

75%           1.000000                1.000000                0.000000

max           1.000000                1.000000                1.000000


       domain_registration_length    domain_age   web_traffic
dns_record  \
count                11430.000000  11430.000000  1.143000e+04
11430.000000
mean                   492.532196   4062.543745  8.567566e+05
0.020122
std                    814.769415   3107.784600  1.995606e+06
0.140425
min                     -1.000000    -12.000000  0.000000e+00
0.000000
25%                     84.000000    972.250000  0.000000e+00
0.000000
50%                    242.000000   3993.000000  1.651000e+03
0.000000
75%                    449.000000   7026.750000  3.738455e+05
0.000000
max                  29829.000000  12874.000000  1.076799e+07
1.000000

       google_index     page_rank
count  11430.000000  11430.000000
mean       0.533946      3.185739
std        0.498868      2.536955
min        0.000000      0.000000
25%        0.000000      1.000000
50%        1.000000      3.000000
75%        1.000000      5.000000
max        1.000000     10.000000
```

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.

```
numerical_columns = df.select_dtypes(exclude= 'object')
numerical_columns
```

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

       nb_and  nb_or  nb_eq  nb_underscore  nb_tilde  nb_percent
nb_slash  \
0           0      0      0              0         0           0
3
1           0      0      0              0         0           0
5
2           2      0      3              2         0           0
5
3           0      0      0              0         0           0
2
4           0      0      0              0         0           0
5
...       ...    ...    ...            ...       ...         ...
...
11425       0      0      0              0         0           0
4
11426       0      0      1              0         0           1
5
11427       0      0      1              1         0           0
5
11428       0      0      0              0         0           0
3
11429       9      0      9             18         0          23
4
```

```
       nb_star  nb_colon  nb_comma  nb_semicolumn  nb_dollar  nb_space
\
0            0         1         0              0          0         0

1            0         1         0              0          0         0

2            0         1         0              0          0         0

3            0         1         0              0          0         0

4            0         1         0              0          0         0

...        ...       ...       ...            ...        ...       ...

11425        0         1         0              0          0         0

11426        0         1         0              0          0         1

11427        0         1         0              0          0         0

11428        0         1         0              0          0         0

11429        0         1         0              9          0         0


       nb_www  nb_com  nb_dslash  http_in_path  https_token
ratio_digits_url  \
0           1       0          0             0            1
0.000000
1           0       0          0             0            1
0.220779
2           0       1          0             0            0
0.150794
3           0       0          0             0            1
0.000000
4           1       0          0             0            1
0.000000
...       ...     ...        ...           ...          ...
...
11425       1       0          0             0            1
0.000000
11426       1       1          0             0            1
0.023810
11427       1       0          0             0            0
0.142857
11428       1       0          0             0            1
0.000000
11429       1       0          0             4            1
0.085954
```

```
       ratio_digits_host  punycode  port  tld_in_path
tld_in_subdomain  \
0               0.000000         0     0            0
0
1               0.000000         0     0            0
0
2               0.000000         0     0            0
1
3               0.000000         0     0            0
0
4               0.000000         0     0            0
0
...                  ...       ...   ...          ...                ..
.
11425           0.000000         0     0            0
0
11426           0.000000         0     0            1
0
11427           0.000000         0     0            0
0
11428           0.000000         0     0            0
0
11429           0.785714         0     0            1
1

       abnormal_subdomain  nb_subdomains  prefix_suffix  random_domain
\
0                       0              3              0              0

1                       0              1              0              0

2                       0              3              1              0

3                       0              2              0              0

4                       0              2              0              0

...                   ...            ...            ...            ...

11425                   0              2              0              0

11426                   0              3              0              0

11427                   0              2              0              0

11428                   0              2              0              0

11429                   1              3              0              0


       shortening_service  path_extension  nb_redirection  \
```

```
0                        0                0                0
1                        0                0                1
2                        0                0                1
3                        0                0                1
4                        0                0                1
...                    ...              ...              ...
11425                    0                0                1
11426                    0                0                1
11427                    0                0                0
11428                    0                0                0
11429                    0                0                1

       nb_external_redirection  length_words_raw  char_repeat  \
0                            0                 4            4
1                            0                 4            4
2                            0                12            2
3                            0                 1            0
4                            0                 6            3
...                        ...               ...          ...
11425                        0                 4            4
11426                        0                12            3
11427                        0                13            5
11428                        0                 2            3
11429                        1                90            8

       shortest_words_raw  shortest_word_host  shortest_word_path  \
0                       3                   3                   3
1                       2                  19                   2
2                       2                   3                   2
3                       5                   5                   0
4                       3                   3                   4
...                   ...                 ...                 ...
11425                   3                   3                   8
11426                   3                   3                   3
11427                   1                   3                   1
11428                   3                   3                   0
11429                   1                   2                   1

       longest_words_raw  longest_word_host  longest_word_path
avg_words_raw  \
0                     11                 11                   6
5.750000
1                     32                 19                  32
15.750000
2                     17                 13                  17
8.250000
3                      5                  5                   0
5.000000
4                     11                  7                  11
6.333333
```

```
...                   ...                 ...                 ...
...
11425                  11                   9                  11
7.750000
11426                  10                  10                   8
5.166667
11427                  15                   8                  15
6.153846
11428                  22                  22                   0
12.500000
11429                  12                   3                  12
4.377778

       avg_word_host  avg_word_path  phish_hints  domain_in_brand  \
0               7.00       4.500000            0                0
1              19.00      14.666667            0                0
2               8.40       8.142857            0                0
3               5.00       0.000000            0                0
4               5.00       7.000000            0                0
...              ...            ...          ...              ...
11425           6.00       9.500000            0                0
11426           6.50       4.900000            0                0
11427           5.50       6.272727            0                1
11428          12.50       0.000000            0                0
11429           2.75       4.453488            3                0

       brand_in_subdomain  brand_in_path  suspecious_tld
statistical_report  \
0                       0              0               0
0
1                       0              0               0
0
2                       0              0               0
0
3                       0              0               0
0
4                       0              0               0
0
...                   ...            ...             ...
...
11425                   0              0               0
0
11426                   0              0               0
0
11427                   0              0               0
0
11428                   0              0               0
0
11429                   0              1               0
```

```
2

       nb_hyperlinks  ratio_intHyperlinks  ratio_extHyperlinks  \
0                 17             0.529412             0.470588
1                 30             0.966667             0.033333
2                  4             1.000000             0.000000
3                149             0.973154             0.026846
4                102             0.470588             0.529412
...              ...                  ...                  ...
11425            199             0.884422             0.115578
11426              3             1.000000             0.000000
11427             68             0.470588             0.529412
11428             32             0.375000             0.625000
11429             21             0.428571             0.571429

       ratio_nullHyperlinks  nb_extCSS  ratio_intRedirection  \
0                         0          0                     0
1                         0          0                     0
2                         0          0                     0
3                         0          0                     0
4                         0          0                     0
...                     ...        ...                   ...
11425                     0          0                     0
11426                     0          0                     0
11427                     0          5                     0
11428                     0          1                     0
11429                     0          3                     0

       ratio_extRedirection  ratio_intErrors  ratio_extErrors
login_form  \
0                  0.875000                0         0.500000
0
1                  0.000000                0         0.000000
0
2                  0.000000                0         0.000000
0
3                  0.250000                0         0.250000
0
4                  0.537037                0         0.018519
1
...                     ...              ...              ...
...
11425              0.043478                0         0.173913
0
11426              0.000000                0         0.000000
0
11427              0.000000                0         0.000000
0
11428              0.050000                0         0.050000
0
```

```
11429              0.000000                0         0.083333
1

       external_favicon  links_in_tags  submit_email
ratio_intMedia  \
0                     0      80.000000             0      100.000000

1                     0     100.000000             0       80.000000

2                     0     100.000000             0        0.000000

3                     0     100.000000             0       96.428571

4                     0      76.470588             0        0.000000

...                 ...            ...           ...             ...

11425                 1      80.000000             0       21.052632

11426                 0     100.000000             0        0.000000

11427                 1       6.250000             0        0.000000

11428                 1      16.666667             0        0.000000

11429                 1       0.000000             0        0.000000


       ratio_extMedia  sfh  iframe  popup_window  safe_anchor
onmouseover  \
0            0.000000    0       0             0     0.000000
0
1           20.000000    0       0             0   100.000000
0
2            0.000000    0       0             0   100.000000
0
3            3.571429    0       0             0    62.500000
0
4          100.000000    0       0             0     0.000000
0
...               ...  ...     ...           ...          ...
...
11425       78.947368    0       0             0     0.000000
0
11426        0.000000    0       0             0     0.000000
0
11427        0.000000    0       0             0    80.000000
0
11428      100.000000    0       0             0     0.000000
0
```

```
11429        0.000000    0       0             0    33.333333
0

       right_clic  empty_title  domain_in_title  domain_with_copyright
\
0               0            0                0                      1

1               0            0                1                      0

2               0            0                1                      0

3               0            0                1                      0

4               0            0                0                      1

...           ...          ...              ...                    ...

11425           0            0                0                      0

11426           0            0                1                      0

11427           0            0                0                      0

11428           0            0                1                      0

11429           0            0                1                      1


       whois_registered_domain  domain_registration_length  domain_age
\
0                            0                          45          -1

1                            0                          77        5767

2                            0                          14        4004

3                            0                          62          -1

4                            0                         224        8175

...                        ...                         ...         ...

11425                        0                         448        5396

11426                        0                         211        6728

11427                        0                        2809        8515

11428                        0                          85        2836

11429                        1                           0          -1
```

```
       web_traffic  dns_record  google_index  page_rank
0                0           1             1          4
1                0           0             1          2
2          5828815           0             1          0
3           107721           0             0          3
4             8725           0             0          6
...            ...         ...           ...        ...
11425         3980           0             0          6
11426            0           0             1          0
11427            8           0             1         10
11428      2455493           0             0          4
11429            0           1             1          0

[11430 rows x 87 columns]
```

```
# 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
```

```
                      Feature  Minimum       Maximum            Mean
Mode  \
0                  length_url     12.0  1.641000e+03       61.126684
26.0
1              length_hostname      4.0  2.140000e+02       21.090289
```

```
16.0
2                          ip      0.0  1.000000e+00        0.150569
0.0
3                    nb_dots      1.0  2.400000e+01        2.480752
2.0
4                 nb_hyphens      0.0  4.300000e+01        0.997550
0.0
5                    nb_at      0.0  4.000000e+00        0.022222
0.0
6                    nb_qm      0.0  3.000000e+00        0.141207
0.0
7                   nb_and      0.0  1.900000e+01        0.162292
0.0
8                    nb_or      0.0  0.000000e+00        0.000000
0.0
9                    nb_eq      0.0  1.900000e+01        0.293176
0.0
10              nb_underscore      0.0  1.800000e+01        0.322660
0.0
11                  nb_tilde      0.0  1.000000e+00        0.006649
0.0
12                nb_percent      0.0  9.600000e+01        0.123097
0.0
13                  nb_slash      2.0  3.300000e+01        4.289589
3.0
14                   nb_star      0.0  1.000000e+00        0.000700
0.0
15                  nb_colon      1.0  7.000000e+00        1.027909
1.0
16                  nb_comma      0.0  4.000000e+00        0.004024
0.0
17              nb_semicolumn      0.0  2.000000e+01        0.062292
0.0
18                 nb_dollar      0.0  6.000000e+00        0.001925
0.0
19                  nb_space      0.0  1.800000e+01        0.034821
0.0
20                    nb_www      0.0  2.000000e+00        0.448469
0.0
21                    nb_com      0.0  6.000000e+00        0.127997
0.0
22                 nb_dslash      0.0  1.000000e+00        0.006562
0.0
23               http_in_path      0.0  4.000000e+00        0.016710
0.0
24                https_token      0.0  1.000000e+00        0.610936
1.0
25           ratio_digits_url      0.0  7.238806e-01        0.053137
0.0
```

26	ratio_digits_host	0.0	8.000000e-01	0.025024	0.0
27	punycode	0.0	1.000000e+00	0.000350	0.0
28	port	0.0	1.000000e+00	0.002362	0.0
29	tld_in_path	0.0	1.000000e+00	0.065617	0.0
30	tld_in_subdomain	0.0	1.000000e+00	0.050131	0.0
31	abnormal_subdomain	0.0	1.000000e+00	0.021610	0.0
32	nb_subdomains	1.0	3.000000e+00	2.231671	2.0
33	prefix_suffix	0.0	1.000000e+00	0.202450	0.0
34	random_domain	0.0	1.000000e+00	0.083290	0.0
35	shortening_service	0.0	1.000000e+00	0.123447	0.0
36	path_extension	0.0	1.000000e+00	0.000175	0.0
37	nb_redirection	0.0	6.000000e+00	0.498250	0.0
38	nb_external_redirection	0.0	1.000000e+00	0.003150	0.0
39	length_words_raw	1.0	1.060000e+02	6.232808	2.0
40	char_repeat	0.0	1.460000e+02	2.927472	3.0
41	shortest_words_raw	1.0	3.100000e+01	3.127297	3.0
42	shortest_word_host	1.0	3.900000e+01	5.019773	3.0
43	shortest_word_path	0.0	4.000000e+01	2.398950	0.0
44	longest_words_raw	2.0	8.290000e+02	15.393876	9.0
45	longest_word_host	1.0	6.200000e+01	10.467979	9.0
46	longest_word_path	0.0	8.290000e+02	10.561505	0.0
47	avg_words_raw	2.0	1.282500e+02	7.258882	6.0
48	avg_word_host	1.0	3.900000e+01	7.678075	5.0
49	avg_word_path	0.0	2.500000e+02	5.092425	0.0
50	phish_hints	0.0	1.000000e+01	0.327734	

```
0.0
51              domain_in_brand      0.0  1.000000e+00       0.104199
0.0
52           brand_in_subdomain      0.0  1.000000e+00       0.004112
0.0
53                brand_in_path      0.0  1.000000e+00       0.004899
0.0
54              suspecious_tld      0.0  1.000000e+00       0.017935
0.0
55           statistical_report      0.0  2.000000e+00       0.059755
0.0
56                nb_hyperlinks      0.0  4.659000e+03      87.189764
0.0
57          ratio_intHyperlinks      0.0  1.000000e+00       0.602457
0.0
58          ratio_extHyperlinks      0.0  1.000000e+00       0.276720
0.0
59         ratio_nullHyperlinks      0.0  0.000000e+00       0.000000
0.0
60                   nb_extCSS      0.0  1.240000e+02       0.784864
0.0
61         ratio_intRedirection      0.0  0.000000e+00       0.000000
0.0
62         ratio_extRedirection      0.0  2.000000e+00       0.158926
0.0
63              ratio_intErrors      0.0  0.000000e+00       0.000000
0.0
64              ratio_extErrors      0.0  1.000000e+00       0.062469
0.0
65                   login_form      0.0  1.000000e+00       0.063605
0.0
66             external_favicon      0.0  1.000000e+00       0.442170
0.0
67                links_in_tags      0.0  1.000000e+02      51.978211
0.0
68                 submit_email      0.0  0.000000e+00       0.000000
0.0
69               ratio_intMedia      0.0  1.000000e+02      42.870444
0.0
70               ratio_extMedia      0.0  1.000000e+02      23.236293
0.0
71                          sfh      0.0  0.000000e+00       0.000000
0.0
72                       iframe      0.0  1.000000e+00       0.001312
0.0
73                 popup_window      0.0  1.000000e+00       0.006037
0.0
74                  safe_anchor      0.0  1.000000e+02      37.063922
0.0
```

```
75                 onmouseover      0.0  1.000000e+00       0.001137
0.0
76                  right_clic      0.0  1.000000e+00       0.001400
0.0
77                 empty_title      0.0  1.000000e+00       0.124759
0.0
78             domain_in_title      0.0  1.000000e+00       0.775853
1.0
79       domain_with_copyright      0.0  1.000000e+00       0.439545
0.0
80     whois_registered_domain      0.0  1.000000e+00       0.072878
0.0
81  domain_registration_length     -1.0  2.982900e+04     492.532196
0.0
82                  domain_age    -12.0  1.287400e+04    4062.543745
-1.0
83                web_traffic      0.0  1.076799e+07  856756.643307
0.0
84                  dns_record      0.0  1.000000e+00       0.020122
0.0
85                google_index      0.0  1.000000e+00       0.533946
1.0
86                   page_rank      0.0  1.000000e+01       3.185739
0.0

           25%            75%            IQR  Standard Deviation
Skewness  \
0    33.000000      71.000000      38.000000        5.529732e+01
8.085190
1    15.000000      24.000000       9.000000        1.077717e+01
5.160078
2     0.000000       0.000000       0.000000        3.576436e-01
1.954418
3     2.000000       3.000000       1.000000        1.369686e+00
5.718117
4     0.000000       1.000000       1.000000        2.087087e+00
4.695239
5     0.000000       0.000000       0.000000        1.554999e-01
8.272893
6     0.000000       0.000000       0.000000        3.644558e-01
2.488737
7     0.000000       0.000000       0.000000        8.213374e-01
9.725295
8     0.000000       0.000000       0.000000        0.000000e+00
0.000000
9     0.000000       0.000000       0.000000        9.983172e-01
6.530036
10    0.000000       0.000000       0.000000        1.093336e+00
7.265925
```

```
11    0.000000       0.000000       0.000000        8.127444e-02
12.142493
12    0.000000       0.000000       0.000000        1.466450e+00
35.424119
13    3.000000       5.000000       2.000000        1.882251e+00
2.731312
14    0.000000       0.000000       0.000000        2.644776e-02
37.764070
15    1.000000       1.000000       0.000000        2.403255e-01
11.383138
16    0.000000       0.000000       0.000000        1.032395e-01
31.694875
17    0.000000       0.000000       0.000000        5.981896e-01
16.160130
18    0.000000       0.000000       0.000000        7.711078e-02
55.657186
19    0.000000       0.000000       0.000000        3.755757e-01
25.424087
20    0.000000       1.000000       1.000000        5.019124e-01
0.261215
21    0.000000       0.000000       0.000000        3.790079e-01
3.778379
22    0.000000       0.000000       0.000000        8.074153e-02
12.224804
23    0.000000       0.000000       0.000000        1.693581e-01
14.818119
24    0.000000       1.000000       1.000000        4.875592e-01  -
0.455147
25    0.000000       0.079365       0.079365        8.936273e-02
2.185535
26    0.000000       0.000000       0.000000        9.342200e-02
5.591198
27    0.000000       0.000000       0.000000        1.870466e-02
53.434537
28    0.000000       0.000000       0.000000        4.854720e-02
20.504781
29    0.000000       0.000000       0.000000        2.476219e-01
3.509054
30    0.000000       0.000000       0.000000        2.182252e-01
4.123698
31    0.000000       0.000000       0.000000        1.454121e-01
6.580939
32    2.000000       3.000000       1.000000        6.370688e-01  -
0.243003
33    0.000000       0.000000       0.000000        4.018432e-01
1.481187
34    0.000000       0.000000       0.000000        2.763315e-01
3.016546
35    0.000000       0.000000       0.000000        3.289641e-01
```

```
2.289726
36    0.000000       0.000000       0.000000        1.322735e-02
75.587696
37    0.000000       1.000000       1.000000        6.919070e-01
1.568452
38    0.000000       0.000000       0.000000        5.603535e-02
17.736565
39    2.000000       8.000000       6.000000        5.572355e+00
5.367350
40    1.000000       4.000000       3.000000        4.768936e+00
15.756781
41    2.000000       3.000000       1.000000        2.211571e+00
3.156658
42    3.000000       6.000000       3.000000        3.941580e+00
2.267936
43    0.000000       3.000000       3.000000        2.997809e+00
4.687604
44    9.000000      16.000000       7.000000        2.208364e+01
13.531024
45    7.000000      13.000000       6.000000        4.932015e+00
1.631076
46    0.000000      11.000000      11.000000        2.307788e+01
12.395730
47    5.250000       8.000000       2.750000        4.145827e+00
9.548722
48    5.250000       9.000000       3.750000        3.578435e+00
1.746080
49    0.000000       6.714286       6.714286        7.147050e+00
13.446741
50    0.000000       0.000000       0.000000        8.426004e-01
3.216484
51    0.000000       0.000000       0.000000        3.055325e-01
2.591343
52    0.000000       0.000000       0.000000        6.399559e-02
15.500280
53    0.000000       0.000000       0.000000        6.982700e-02
14.183260
54    0.000000       0.000000       0.000000        1.327220e-01
7.265550
55    0.000000       0.000000       0.000000        3.312662e-01
5.516960
56    9.000000     101.000000      92.000000        1.667583e+02
7.675061
57    0.224991       0.944767       0.719776        3.764745e-01  -
0.528009
58    0.000000       0.474840       0.474840        3.199583e-01
1.008455
59    0.000000       0.000000       0.000000        0.000000e+00
0.000000
```

```
60    0.000000       1.000000       1.000000        2.758802e+00
23.495479
61    0.000000       0.000000       0.000000        0.000000e+00
0.000000
62    0.000000       0.230769       0.230769        2.664370e-01
2.296810
63    0.000000       0.000000       0.000000        0.000000e+00
0.000000
64    0.000000       0.034483       0.034483        1.562087e-01
3.510798
65    0.000000       0.000000       0.000000        2.440578e-01
3.576790
66    0.000000       1.000000       1.000000        4.966661e-01
0.232915
67    0.000000      98.061004      98.061004        4.152314e+01  -
0.150770
68    0.000000       0.000000       0.000000        0.000000e+00
0.000000
69    0.000000     100.000000     100.000000        4.624990e+01
0.275813
70    0.000000      33.333333      33.333333        3.838658e+01
1.265615
71    0.000000       0.000000       0.000000        0.000000e+00
0.000000
72    0.000000       0.000000       0.000000        3.620398e-02
27.553595
73    0.000000       0.000000       0.000000        7.746501e-02
12.755428
74    0.000000      75.000000      75.000000        3.907339e+01
0.513075
75    0.000000       0.000000       0.000000        3.370703e-02
29.605100
76    0.000000       0.000000       0.000000        3.738968e-02
26.675141
77    0.000000       0.000000       0.000000        3.304604e-01
2.271415
78    1.000000       1.000000       0.000000        4.170376e-01  -
1.323148
79    0.000000       1.000000       1.000000        4.963535e-01
0.243639
80    0.000000       0.000000       0.000000        2.599482e-01
3.286781
81   84.000000     449.000000     365.000000        8.147694e+02
9.819607
82  972.250000    7026.750000    6054.500000        3.107785e+03
0.164187
83    0.000000  373845.500000  373845.500000        1.995606e+06
2.779269
84    0.000000       0.000000       0.000000        1.404254e-01
```

```
6.835821
85    0.000000        1.000000        1.000000         4.988682e-01  -
0.136115
86    1.000000        5.000000        4.000000         2.536955e+00
0.446031

       Kurtosis
0    144.196391
1     69.829931
2      1.820067
3     66.155843
4     40.696686
5     95.457038
6      6.060591
7    139.140959
8      0.000000
9     70.909580
10    83.018490
11   145.465581
12  1872.381213
13    18.302403
14  1424.374235
15   166.331068
16  1112.515623
17   339.259665
18  3695.607690
19   997.130918
20    -1.792086
21    23.443303
22   147.471633
23   284.259172
24    -1.793155
25     5.343812
26    36.596651
27  2853.749081
28   418.519283
29    10.315261
30    15.007510
31    41.315987
32    -0.667307
33     0.193948
34     7.100790
35     3.243414
36  5712.499300
37     3.524335
38   312.640431
39    60.762427
40   393.756615
41    15.725890
```

```
42     6.957438
43    38.278180
44   294.780580
45     7.201450
46   256.859316
47   189.456143
48     5.933412
49   336.108296
50    12.518737
51     4.715885
52   238.300374
53   199.199720
54    50.797103
55    28.928927
56   117.966663
57    -1.297591
58    -0.320285
59     0.000000
60   887.252101
61     0.000000
62     6.639213
63     0.000000
64    13.627934
65    10.795317
66    -1.946091
67    -1.667946
68     0.000000
69    -1.821300
70    -0.195738
71     0.000000
72   757.333090
73   160.729074
74    -1.349057
75   874.614982
76   709.687338
77     3.159878
78    -0.249322
79    -1.940980
80     8.804472
81   294.677934
82    -1.097305
83     7.306645
84    44.736280
85    -1.981820
86    -0.386315
```

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).

```
# 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.amazon.co.jp%2F%3Fref_
%3Dnav_em_hd_re_signin&amp;openid.identity=http%3A%2F
%2Fspecs.openid.net%2Fauth
%2F2.0%2Fidentifier_select&amp;openid.assoc_handle=jpflex&amp;openid.m
ode=checkid_setup&amp;key=a@b.c&amp;openid.claimed_id=http%3A%2F
%2Fspecs.openid.net%2Fauth
%2F2.0%2Fidentifier_select&amp;openid.ns=http%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
```

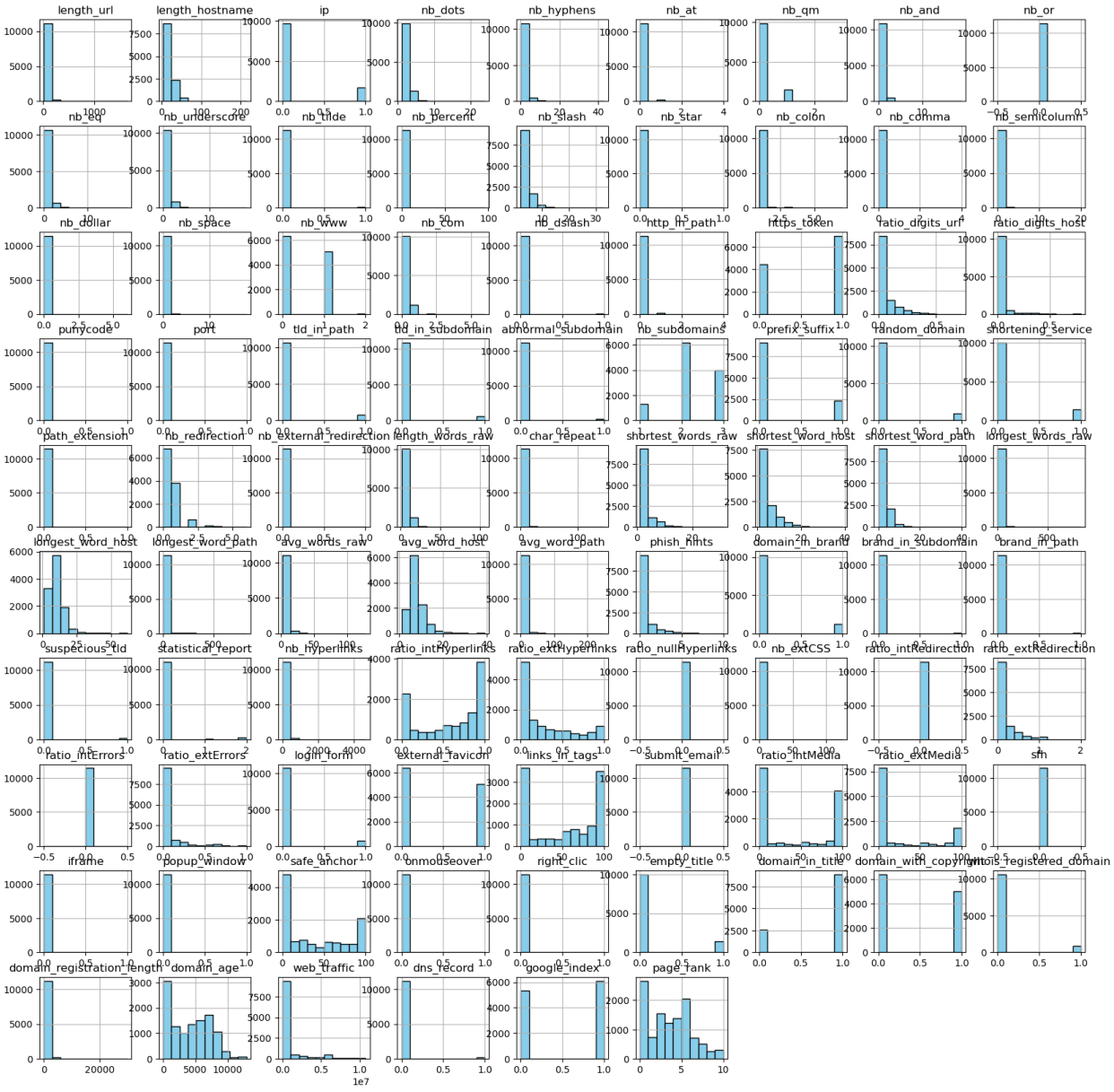
The target label is balanced — There is no need to use SMOTE techniques to Blanace the Target columm.

```
df['status'].mode()

0    legitimate
1      phishing
Name: status, dtype: object

df['url'].mode()

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.

```
# 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()
```

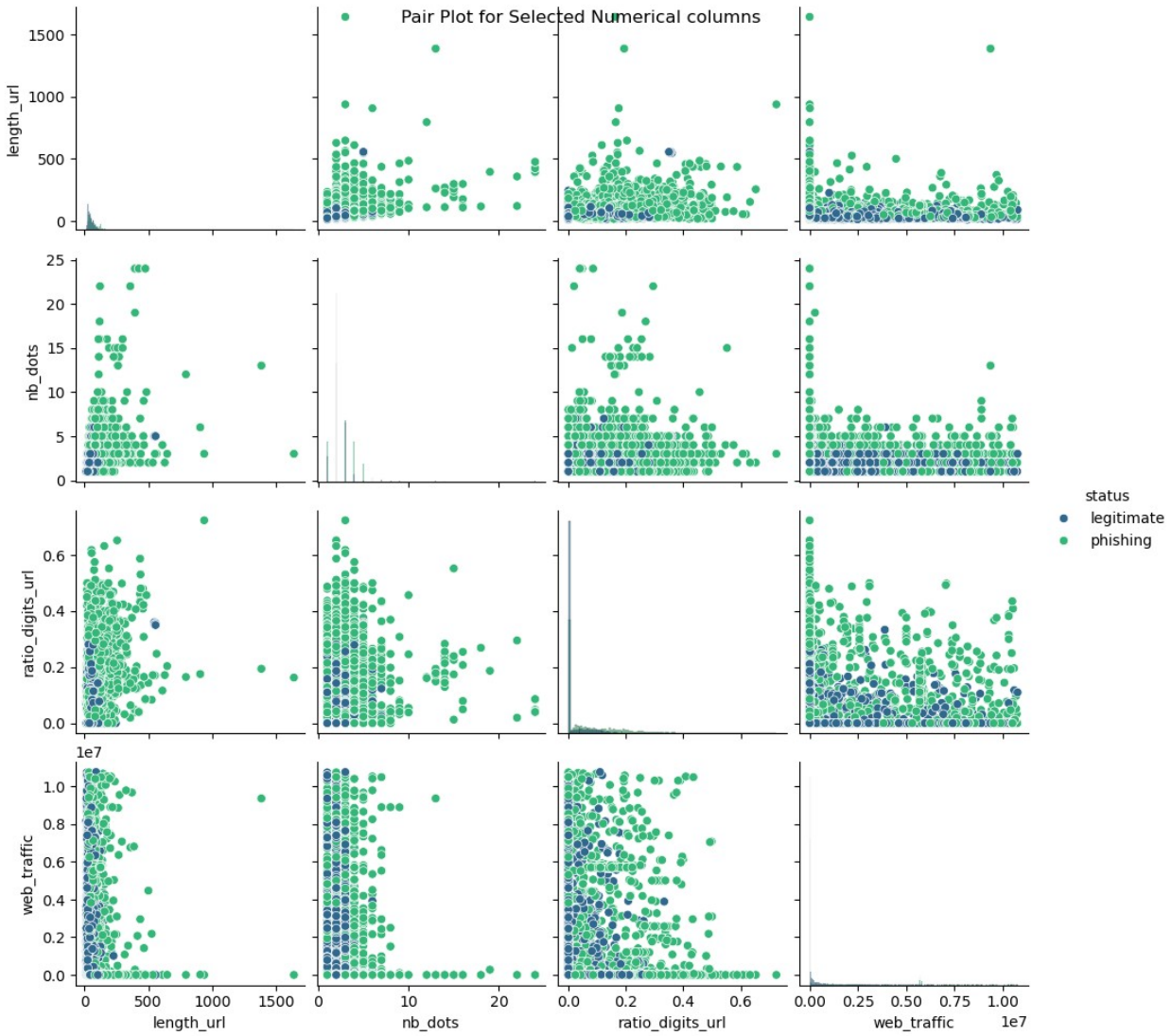


Pair Plot

- 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.

```
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.

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

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

```
# Using Label Encoding in Url column
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

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

```
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.