Author: Saikumar Mehtre

In [1]:

```
#Import Data Manupulation Library
import numpy as np
import pandas as pd

from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.decomposition import PCA

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

#Import Filter Warnings Library
import warnings
warnings.filterwarnings("ignore")

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

Loading Dataset

In [3]:

```
#Import Data Using Pandas Function

url="https://raw.githubusercontent.com/Saimehtre18/CodeB_Intership/refs/heads/main/dataset_phishing.csv"
df=pd.read_csv(url)
df.sample(frac=1) # Shuffle the DataFrame rows
df
```

Out[3]:

	url	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_and
0	http://www.crestonwood.com/router.php	37	19	0	3	0	0	0	0
1	http://shadetreetechnology.com/V4/validation/a...	77	23	1	1	0	0	0	0
2	https://support-appleld.com.secureupdate.duila...	126	50	1	4	1	0	1	2
3	http://rgipt.ac.in	18	11	0	2	0	0	0	0
4	http://www.iracing.com/tracks/gateway-motorspo...	55	15	0	2	2	0	0	0
...	...	...	...	...	...	...	...	...	...
11425	http://www.fontspace.com/category/blackletter	45	17	0	2	0	0	0	0
11426	http://www.budgetbots.com/server.php/Server%20...	84	18	0	5	0	1	1	0
11427	https://www.facebook.com/Interactive-Televisio...	105	16	1	2	6	0	1	0
11428	http://www.mypublicdomainpictures.com/	38	30	0	2	0	0	0	0
11429	http://174.139.46.123/ap/signin?openid.pape.ma...	477	14	1	24	0	1	1	9

11430 rows × 89 columns

📊 *Exploratory Data Analysis (EDA)*

Overview of the dataset, including the number of features, types of data (numerical, categorical, etc.), and target variable distribution

In [4]:

```
# Checking Dataset Information
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 [5]:

```
df.shape
```

Out[5]:

```
(11430, 89)
```

In [7]:

```
# Checking Null Value in DataSet
df.isnull().sum()
```

Out[7]:

```
url                0
length_url         0
length_hostname    0
ip                 0
nb_dots            0
                  ..
web_traffic        0
dns_record         0
google_index       0
page_rank          0
status             0
Length: 89, dtype: int64
```

In [10]:

```
# Checking the statistical summary of the dataset
# Checking the correlation between the features
df.describe()
```

Out[10]:

	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	1
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	

8 rows × 87 columns

In [8]:

```

#Summary of numerical features
df.describe()

{"type":"dataframe"}

#Display unique values in categorical features
df.select_dtypes(include=['object']).nunique()
```

Out[8]:

```
url       11429
status        2
dtype: int64
```

Separating numerical and categorical columns

In [22]:

```

Numerical_Data=df.select_dtypes(exclude=['object'])
Categorical_Data=df.select_dtypes(include=['object'])
```

In [23]:

```

Numerical_Data
```

Out[23]:

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

11430 rows × 88 columns

In [24]:

```

Categorical_Data
```

Out[24]:

	url
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...

11430 rows × 1 columns

In [25]:

```
# Checking Descriptive Stats:

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[25]:

	Feature	Minimum	Maximum	Mean	Mode	25%	75%	IQR	Standard Deviation	Skewness	Kurtosis
0	length_url	12.0	1641.0	61.126684	26.0	33.0	71.0	38.0	5.529732e+01	8.085190	144.196391
1	length_hostname	4.0	214.0	21.090289	16.0	15.0	24.0	9.0	1.077717e+01	5.160078	69.829931
2	ip	0.0	1.0	0.150569	0.0	0.0	0.0	0.0	3.576436e-01	1.954418	1.820067
3	nb_dots	1.0	24.0	2.480752	2.0	2.0	3.0	1.0	1.369686e+00	5.718117	66.155843
4	nb_hyphens	0.0	43.0	0.997550	0.0	0.0	1.0	1.0	2.087087e+00	4.695239	40.696686
...	...	...	...	...	...	...	...	...	...	...	...
83	web_traffic	0.0	10767986.0	856756.643307	0.0	0.0	373845.5	373845.5	1.995606e+06	2.779269	7.306645
84	dns_record	0.0	1.0	0.020122	0.0	0.0	0.0	0.0	1.404254e-01	6.835821	44.736280
85	google_index	0.0	1.0	0.533946	1.0	0.0	1.0	1.0	4.988682e-01	-0.136115	-1.981820
86	page_rank	0.0	10.0	3.185739	0.0	1.0	5.0	4.0	2.536955e+00	0.446031	-0.386315
87	status	0.0	1.0	0.500000	0.0	0.0	1.0	1.0	5.000219e-01	0.000000	-2.000350

88 rows × 11 columns

In [26]:

```
df.url.nunique()
```

Out[26]: 11429

In [27]:

```
outlier_label = []
for col in report['Feature']:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    LW = Q1 - 1.5 * IQR
    UW = Q3 + 1.5 * IQR
    outliers = df[(df[col] < LW) | (df[col] > UW)]
    if not outliers.empty:
        outlier_label.append("Has Outliers")
    else:
        outlier_label.append("No Outliers")

report["Outlier Comment"] = outlier_label

report
```

Out[27]:

	Feature	Minimum	Maximum	Mean	Mode	25%	75%	IQR	Standard Deviation	Skewness	Kurtosis	O Com
0	length_url	12.0	1641.0	61.126684	26.0	33.0	71.0	38.0	5.529732e+01	8.085190	144.196391	Ou
1	length_hostname	4.0	214.0	21.090289	16.0	15.0	24.0	9.0	1.077717e+01	5.160078	69.829931	

1	length_hostname	4.0	214.0	21.090289	16.0	15.0	24.0	9.0	1.077717e+01	5.160078	69.829931	Ou
2	ip	0.0	1.0	0.150569	0.0	0.0	0.0	0.0	3.576436e-01	1.954418	1.820067	Ou
3	nb_dots	1.0	24.0	2.480752	2.0	2.0	3.0	1.0	1.369686e+00	5.718117	66.155843	Ou
4	nb_hyphens	0.0	43.0	0.997550	0.0	0.0	1.0	1.0	2.087087e+00	4.695239	40.696686	Ou
...	...	...	...	...	...	...	...	...	...	...	...	
83	web_traffic	0.0	10767986.0	856756.643307	0.0	0.0	373845.5	373845.5	1.995606e+06	2.779269	7.306645	Ou
84	dns_record	0.0	1.0	0.020122	0.0	0.0	0.0	0.0	1.404254e-01	6.835821	44.736280	Ou
85	google_index	0.0	1.0	0.533946	1.0	0.0	1.0	1.0	4.988682e-01	-0.136115	-1.981820	Ou
86	page_rank	0.0	10.0	3.185739	0.0	1.0	5.0	4.0	2.536955e+00	0.446031	-0.386315	Ou
87	status	0.0	1.0	0.500000	0.0	0.0	1.0	1.0	5.000219e-01	0.000000	-2.000350	Ou

88 rows × 12 columns

In [16]:

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

Out[16]:

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

In [17]:

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

Out[17]:

```
0    legitimate
1      phishing
Name: status, dtype: object
```

In [18]:

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

Out[18]:

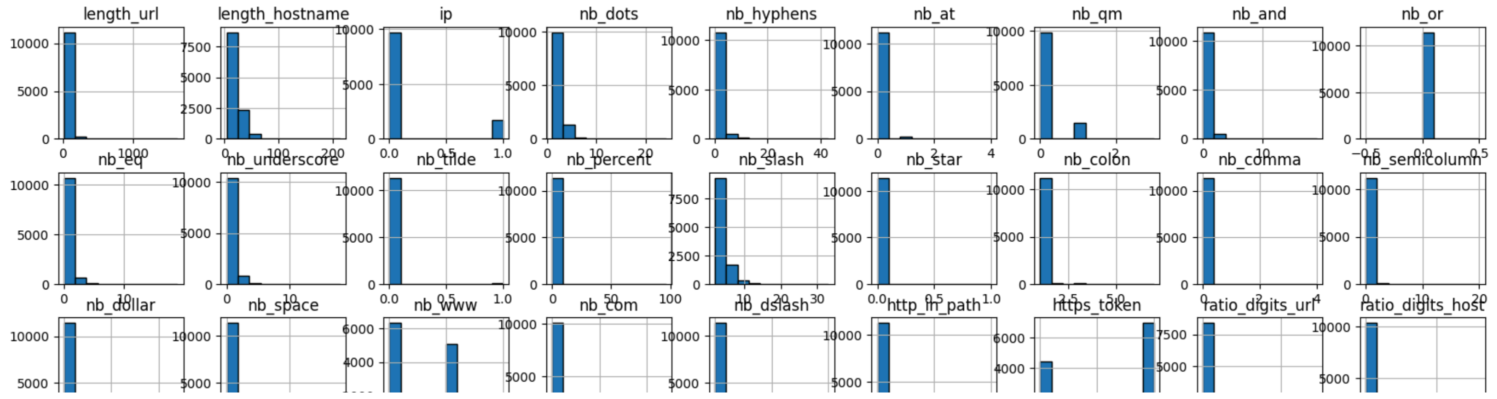
```
0        0
1        1
2        1
3        0
4        0
        ..
11425    0
11426    1
11427    0
11428    0
11429    1
Name: status, Length: 11430, dtype: int64
```

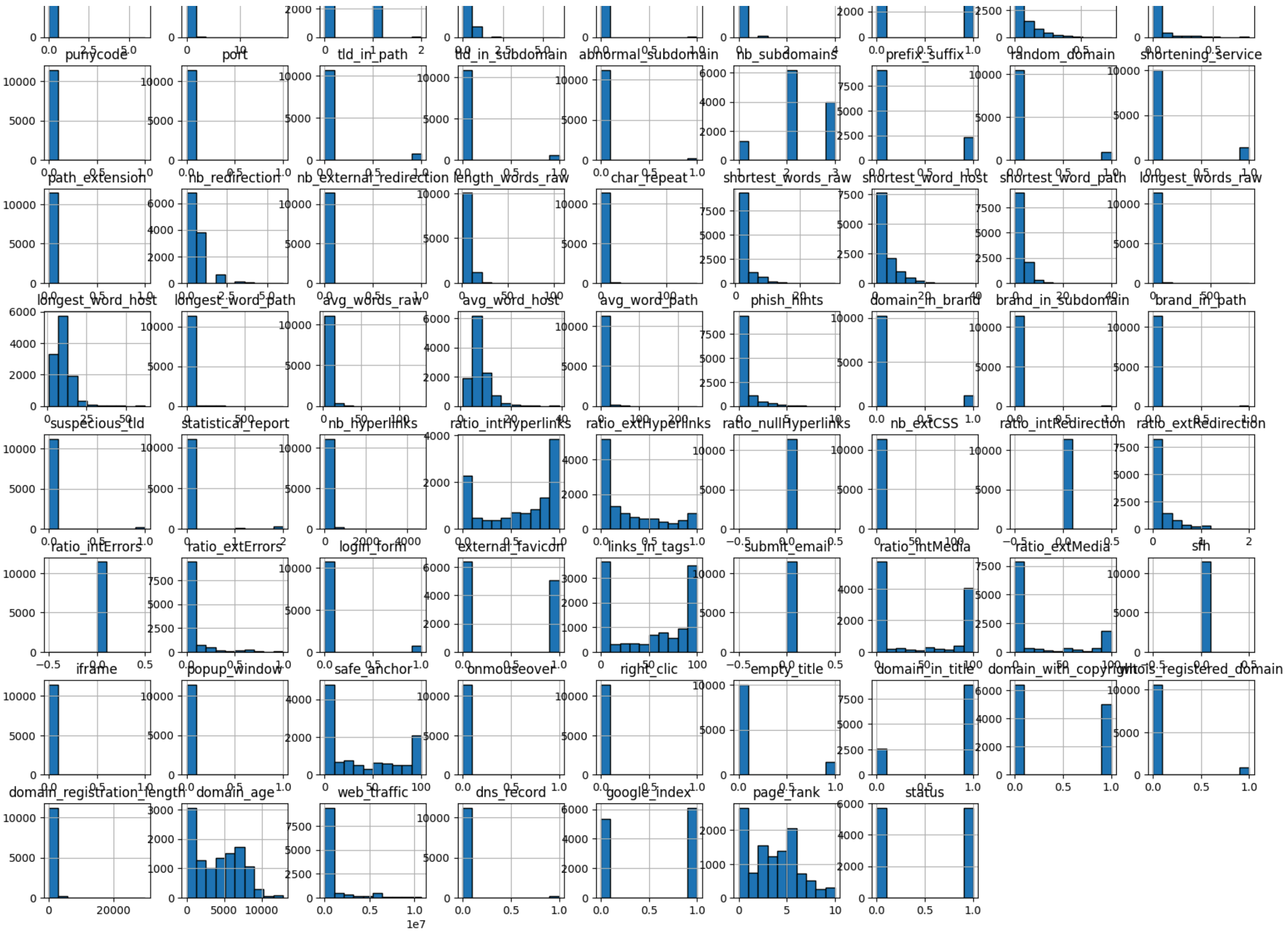
The target variable status was originally categorical, labeled as "phishing" and "legitimate." It was converted into a binary format (1 and 0) for model compatibility.

Histogram

In [28]:

```
#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()
```

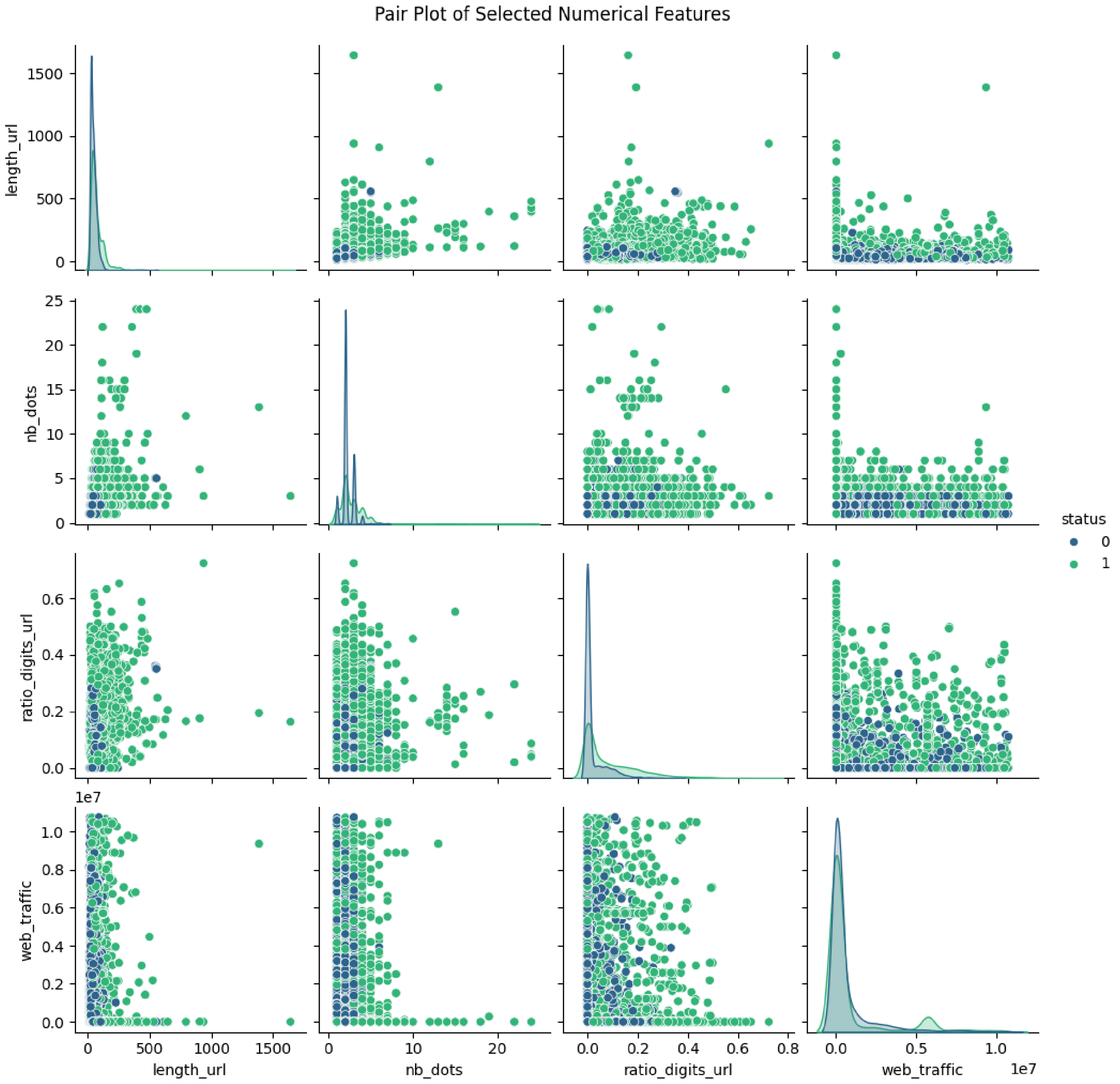




Pair Plot

In [29]:

```
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()
```


Pair Plot of Selected Numerical Features
length_url
nb_dots
ratio_digits_url
web_traffic
1e7
status
1e7