...	...
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	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
```

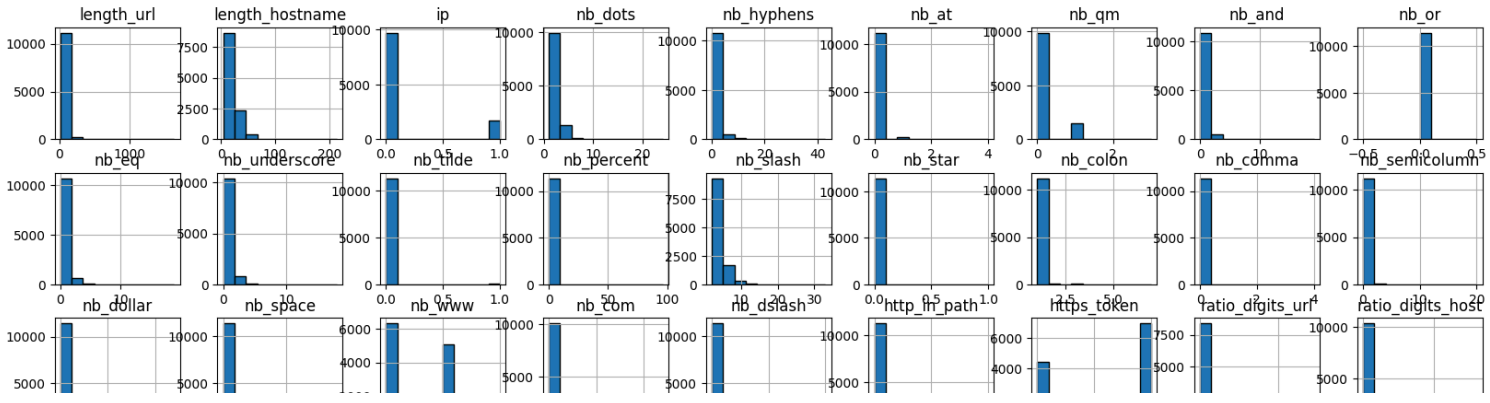
In [18]:
```
# Encoding Target column
df['status']=df['status'].replace({'legitimate':0,'phishing':1})
df['status']
```

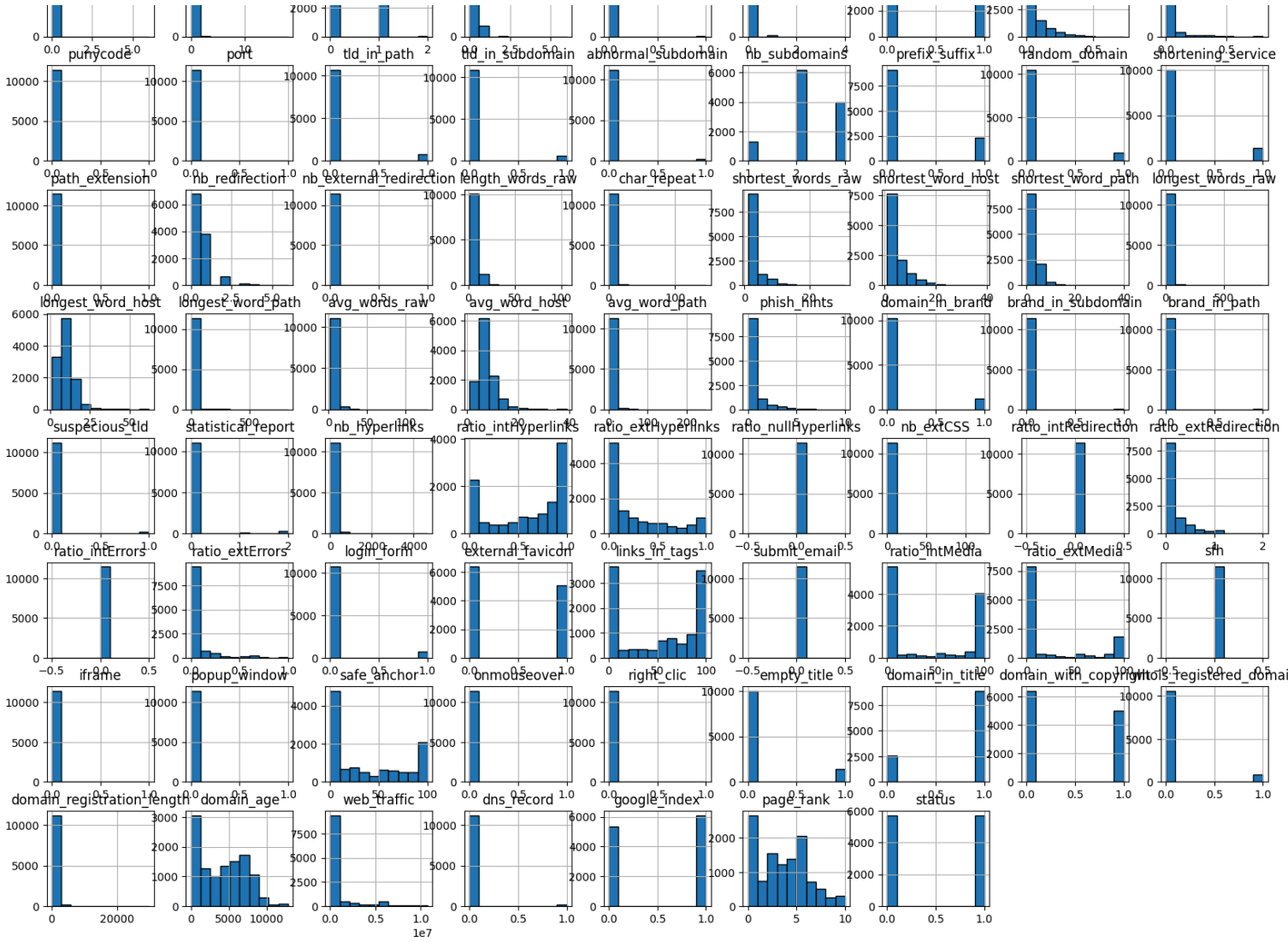
Out[18]:
```
0        0
1        1
2        1
3        0
4        0
        ..
11425    0
11426    1
11427    0
11428    0
11429    1
Name: status, Length: 11430, dtype: int64
```

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()
```