[ ]

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sn
```

[39] ✓ 1s

```
df=pd.read_csv("/content/PhiUSIIL_Phishing_URL_Dataset.csv")
```

[40] ✓ 0s

```
df.head()
```

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContir
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	

```python
df.dropna(inplace=True)
print("Null values after dropping rows:")
print(df.isnull().sum())
```

```
Null values after dropping rows:
FILENAME                    0
URL                         0
URLLength                   0
Domain                      0
DomainLength                0
IsDomainIP                  0
TLD                         0
URLSimilarityIndex          0
CharContinuationRate        0
TLDLegitimateProb           0
URLCharProb                 0
TLDLength                   0
NoOfSubDomain               0
HasObfuscation              0
NoOfObfuscatedChar          0
```

[42]
✓ 0s

```
df.head()
```

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContin
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	

5 rows × 56 columns

[43]
✓ 0s

```
print("Original data type of 'label' column:", df['label'].dtype)
```

```
Original data type of 'label' column: int64
```

[44]
✓ 0s

```
df.head()
```

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContin

[44] ✓ 0s

```
df.head()
```

	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContir
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	

5 rows × 56 columns

[45] ✓ 0s

```
df['label'] = df['label'].astype(int)
print("Data type of 'label' column after conversion:", df['label'].dtype)
```

```
Data type of 'label' column after conversion: int64
```

[46] ✓ 0s

```
df.head()
```

[46]
✓ 0s

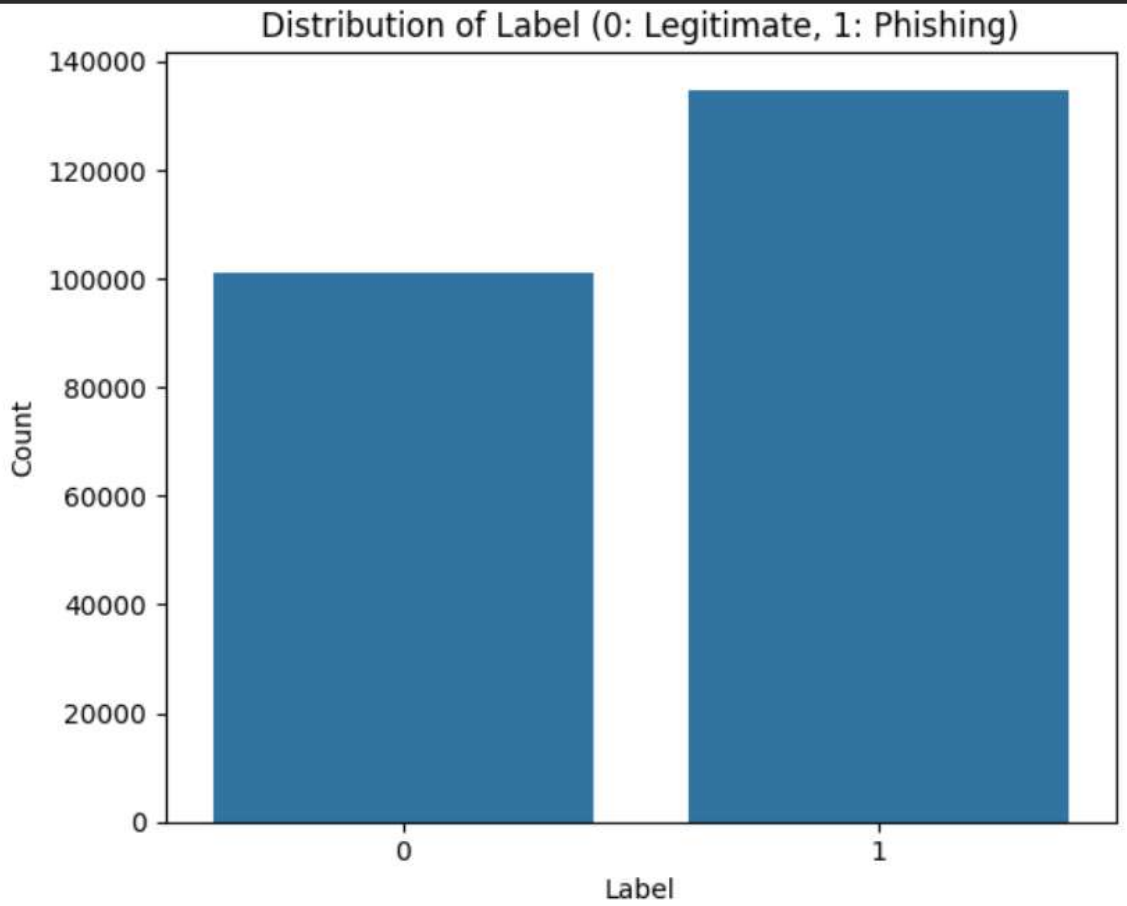
```
df.head()
```

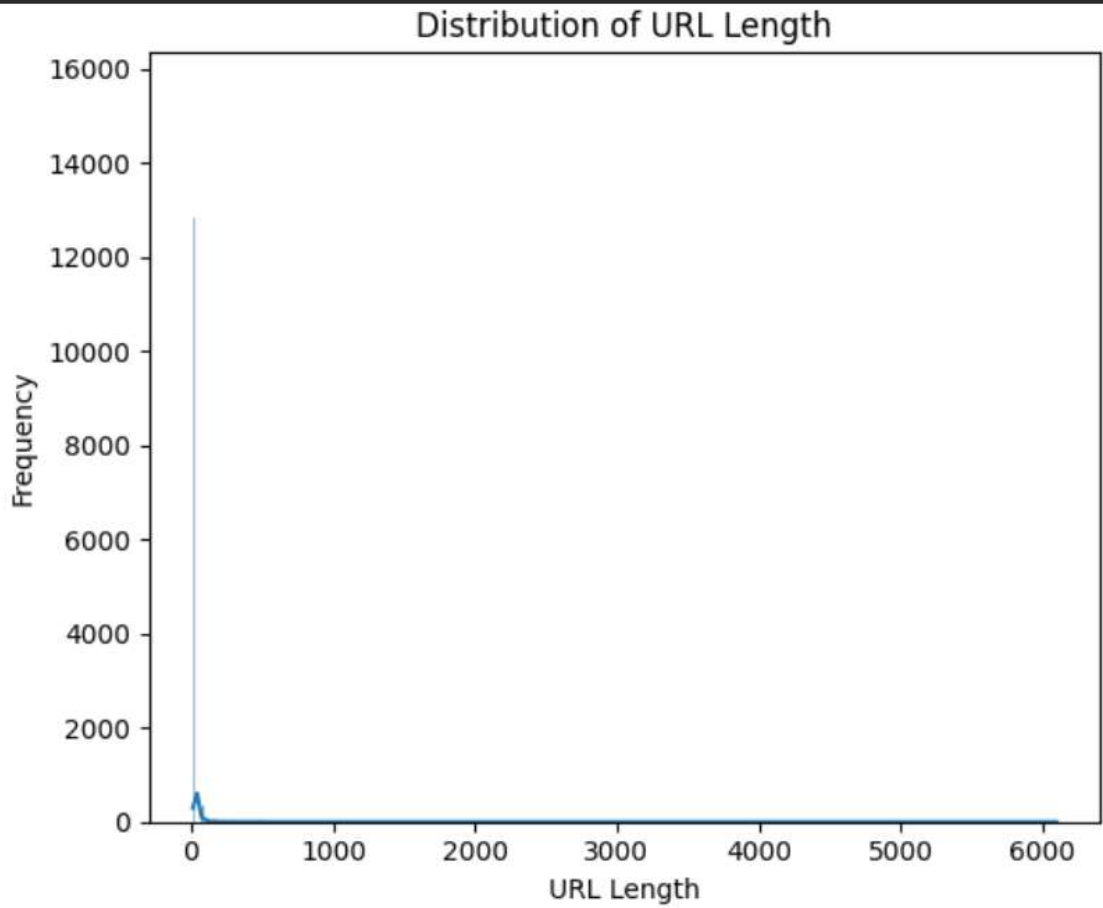
	FILENAME	URL	URLLength	Domain	DomainLength	IsDomainIP	TLD	URLSimilarityIndex	CharContir
0	521848.txt	https://www.southbankmosaics.com	31	www.southbankmosaics.com	24	0	com	100.0	
1	31372.txt	https://www.uni-mainz.de	23	www.uni-mainz.de	16	0	de	100.0	
2	597387.txt	https://www.voicefmradio.co.uk	29	www.voicefmradio.co.uk	22	0	uk	100.0	
3	554095.txt	https://www.sfnmjournal.com	26	www.sfnmjournal.com	19	0	com	100.0	
4	151578.txt	https://www.rewildingargentina.org	33	www.rewildingargentina.org	26	0	org	100.0	

5 rows × 56 columns

```
New features 'NoOfSpecialChars' and 'HasWWW' created.
                                URL                      Domain  \
0    https://www.southbankmosaics.com    www.southbankmosaics.com
1            https://www.uni-mainz.de            www.uni-mainz.de
2      https://www.voicefmradio.co.uk      www.voicefmradio.co.uk
3         https://www.sfnmjournal.com         www.sfnmjournal.com
4  https://www.rewildingargentina.org  www.rewildingargentina.org

   NoOfSpecialChars  HasWWW
0                 0    True
1                 0    True
2                 0    True
3                 0    True
4                 0    True
```


Distribution of Label (0: Legitimate, 1: Phishing)
Count
Label
0
1
0
20000
40000
60000
80000
100000
120000
140000


Distribution of URL Length
Frequency
URL Length
0
2000
4000
6000
8000
10000
12000
14000
16000
0
1000
2000
3000
4000
5000
6000