importing the necessary libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Loading the data at a glance:

In [2]:
```
password_dataset = pd.read_csv('passwords.csv', index_col='rank')
```

Data exploration

In [3]:
```
password_dataset.head(10)
```

Out[3]:

rank	password	category	value	time_unit	offline_crack_sec	rank_alt	strength	font_size
1.0	password	password-related	6.91	years	2.170000e+00	1.0	8.0	11.0
2.0	123456	simple-alphanumeric	18.52	minutes	1.110000e-05	2.0	4.0	8.0
3.0	12345678	simple-alphanumeric	1.29	days	1.110000e-03	3.0	4.0	8.0
4.0	1234	simple-alphanumeric	11.11	seconds	1.110000e-07	4.0	4.0	8.0
5.0	qwerty	simple-alphanumeric	3.72	days	3.210000e-03	5.0	8.0	11.0
6.0	12345	simple-alphanumeric	1.85	minutes	1.110000e-06	6.0	4.0	8.0
7.0	dragon	animal	3.72	days	3.210000e-03	7.0	8.0	11.0
8.0	baseball	sport	6.91	years	2.170000e+00	8.0	4.0	8.0
9.0	football	sport	6.91	years	2.170000e+00	9.0	7.0	11.0
10.0	letmein	password-related	3.19	months	8.350000e-02	10.0	8.0	11.0

```
password_dataset['offline_crack_sec'] = password_dataset['offline_crack_sec'].astype(float).round(2)
```

```
password_dataset.head()
```

rank	password	category	value	time_unit	offline_crack_sec	rank_alt	strength	font_size
1.0	password	password-related	6.91	years	2.17	1.0	8.0	11.0
2.0	123456	simple-alphanumeric	18.52	minutes	0.00	2.0	4.0	8.0
3.0	12345678	simple-alphanumeric	1.29	days	0.00	3.0	4.0	8.0
4.0	1234	simple-alphanumeric	11.11	seconds	0.00	4.0	4.0	8.0
5.0	qwerty	simple-alphanumeric	3.72	days	0.00	5.0	8.0	11.0

```
password_dataset.isnull().sum()
```

```
password            7
category            7
value               7
time_unit           7
offline_crack_sec   7
rank_alt            7
strength            7
font_size           7
dtype: int64
```

```
#Remove rows with missing values

password_dataset.dropna(axis=0, how='any', inplace=True)
```

```
#Check for duplicate data

password_dataset.duplicated().sum()
```

```
0
```

Key Information about the dataset

```
password_dataset.shape
```

```
(500, 8)
```

```
password_dataset.info()
```

```python
#Generate the descriptive statistics of the 'strength' column

strength_stats = password_dataset['strength'].describe()
print(strength_stats)
```

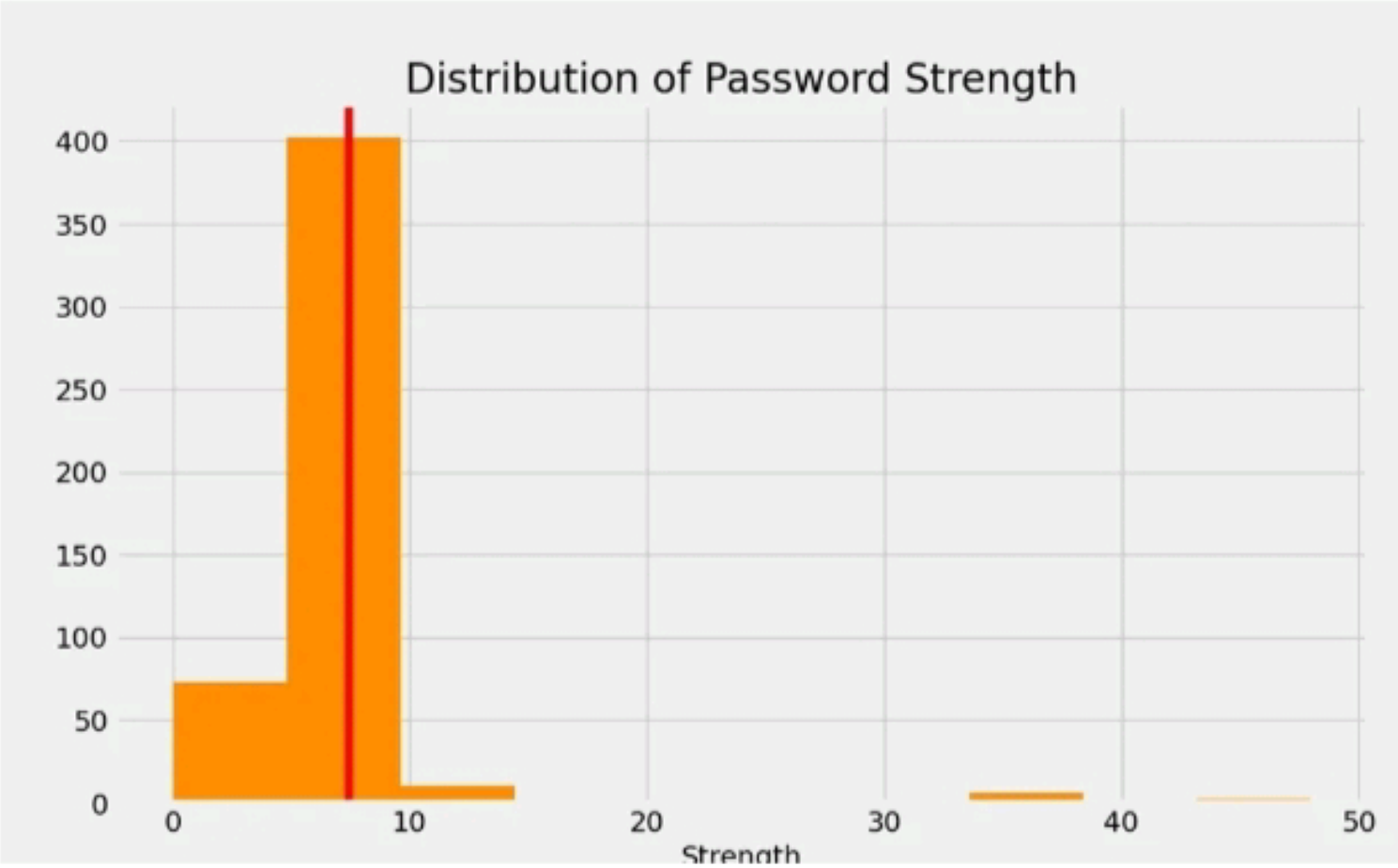
```
count    500.000000
mean       7.432000
std        5.415536
min        0.000000
25%        6.000000
50%        7.000000
75%        8.000000
max       48.000000
Name: strength, dtype: float64
```

Passwords with zero strength:

In [14]:
```
# filter the password with the weakest strength (strength == 0)
weakest_password = password_dataset[password_dataset['strength'] == 0]

# Display some common weakest passwords
common_weakest_password = weakest_password['password']
print('Sample of common weakest_password')
print(common_weakest_password.head())

# Count all weakest passwords
print('.......................')
print("Number of all password with zero strength:")
print(len(weakest_password))
```

```
Sample of common weakest_password
rank
19.0    111111
20.0      2000
46.0    pepper
60.0    666666
77.0      1111
Name: password, dtype: object
.......................
Number of all password with zero strength:
30
```

```
#show a distribution of password strength

plt.figure(figsize=(10,6))
plt.hist(password_dataset['strength'], color='darkorange')
plt.axvline(password_dataset['strength'].mean(), color='red')
plt.style.use('fivethirtyeight')
plt.title('Distribution of Password Strength')
plt.xlabel('Strength')
plt.savefig('strength histogram')
plt.show()

'''
This histogram depicts the spread of passwords' strength.
it's observable that most of the passwords in this dataset are not strong enough.
because they most strength cluster around the central value
'''
```



In [49]:
```
# Detect outliers in the strength of passwords

plt.figure(figsize=(10,6))
plt.boxplot(password_dataset['strength'], vert=False, patch_artist=True)
plt.style.use('ggplot')
plt.title('Extreme Password Strength')
plt.xlabel('Strength')
plt.savefig('strength boxplot')
plt.show()


'''
This is a boxplot to spot outliers in the passwords' strength.
the graph shows that some passwords lie beyond the whiskers.
because their strength are either too high or too low.
Therefore, they are considered potential outliers.
'''
```