65070501037

Paweekorn Soratyathorn

```
# from google.colab import drive
# drive.mount('/content/drive')
```

Web Scraper

```
from bs4 import BeautifulSoup
import requests
```

Require input is the URL.

```
url = 'https://www.britannica.com/topic/list-of-state-capitals-in-the-United-States-2119210'
page = requests.get(url)
soup = BeautifulSoup(page.text, 'html')
print(soup)
```


```
                                                    ,"timezone": "Asia/Bangkok"
                                            ,"bcomId": "-5998384153642313587"
                                                ,"hasAds": true
                                                    ,"testVersion": "D"
                                                        ,"adsTestVersion": "D"
                                                        ,"consumerId": ""
                                                        ,"instId": ""
                                                        ,"consumerUserName": ""
                                                        ,"instUserName": ""
                                                        ,"cognito": null
                                    },
                                    "tvs":{ "r":[25,25,25,25], "a": [25,25,25,25]},
                                    "isLoggedInAsUser": false,
                                    "isPhone": false,
                                    "isDesktop": true,
                                    "logoutUrl": "/auth2/logout",
                                    "selfServiceUrl": "https://myaccount.britannica.com",
                                    "cdnUrl": "https://cdn.britannica.com",
                                    "chatbotApi": "https://www.britannica.com/chat-api",
                            "fetchOffset": 800,
```

```
<link href="https://fonts.googleapis.com/" rel="dns-prefetch"/>
```

Find all tables in the page. In this website, though, there's only one so it's simple.

```
soup.find_all('table')
```

```
[<table> <thead> <tr> <th>state</th> <th>capital</th> <th>population of capital: census</th> <th>population of capital:
estimated</th> </tr> </thead> <tbody> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Alabama-state">Alabama</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Montgomery-Alabama">Montgomery</a></td> <td>(2020) 200,603</td> <td>(2021 est.) 198,665</td>
</tr> <tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Alaska">Alaska</a></td> <td>
<a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Juneau">Juneau</a></td> <td>(2020)
32,255</td> <td>(2021 est.) 31,973</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Arizona-state">Arizona</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Phoenix-Arizona">Phoenix</a></td> <td>(2020) 1,608,139</td> <td>(2021 est.) 1,624,569</td>
</tr> <tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Arkansas-state">Arkansas</a>
</td> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Little-Rock">Little Rock</a></td>
<td>(2020) 202,591</td> <td>(2021 est.) 201,998</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/California-state">California</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Sacramento-California">Sacramento</a></td> <td>(2020) 524,943</td> <td>(2021 est.)
525,041</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Colorado-
state">Colorado</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Denver">Denver</a></td> <td>(2020) 715,522</td> <td>(2021 est.) 711,463</td> </tr> <tr> <td>
<a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Connecticut">Connecticut</a></td> <td><a
class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Hartford-Connecticut">Hartford</a></td> <td>
(2020) 121,054</td> <td>(2021 est.) 120,576</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Delaware-state">Delaware</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Dover-Delaware">Dover</a></td> <td>(2020) 39,403</td> <td>(2021 est.) 38,992</td> </tr> <tr>
<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Florida">Florida</a></td> <td><a
class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Tallahassee">Tallahassee</a></td> <td>(2020)
196,068</td> <td>(2021 est.) 197,102</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Georgia-state">Georgia</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Atlanta-Georgia">Atlanta</a></td> <td>(2020) 498,715</td> <td>(2021 est.) 496,461</td> </tr>
<tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Hawaii-state">Hawaii</a></td> <td>
<a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Honolulu">Honolulu</a></td> <td>(2020)
350,964</td> <td>(2021 est.) 345,510</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Idaho">Idaho</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Boise-Idaho">Boise</a></td> <td>(2020) 235,684</td> <td>(2021 est.) 237,446</td> </tr> <tr>
<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Illinois-state">Illinois</a></td> <td><a
class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Springfield-Illinois">Springfield</a></td> <td>
(2020) 114,394</td> <td>(2021 est.) 113,394</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Indiana-state">Indiana</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Indianapolis-Indiana">Indianapolis</a></td> <td>(2020) 887,642</td> <td>(2021 est.)
882,039</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Iowa-
state">Iowa</a></td> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Des-Moines">Des
Moines</a></td> <td>(2020) 214,133</td> <td>(2021 est.) 212,031</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Kansas">Kansas</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Topeka">Topeka</a></td> <td>(2020) 126,587</td> <td>(2021 est.) 125,963</td> </tr> <tr> <td>
<a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Kentucky">Kentucky</a></td> <td><a
class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Frankfort-Kentucky">Frankfort</a></td> <td>
(2020) 28,602</td> <td>(2021 est.) 28,595</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Louisiana-state">Louisiana</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Baton-Rouge">Baton Rouge</a></td> <td>(2020) 227,470</td> <td>(2021 est.) 222,185</td> </tr>
<tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Maine-state">Maine</a></td> <td><a
class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Augusta-Maine">Augusta</a></td> <td>(2020)
18,899</td> <td>(2021 est.) 18,968</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Maryland-state">Maryland</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Annapolis">Annapolis</a></td> <td>(2020) 40,812</td> <td>(2021 est.) 40,687</td> </tr> <tr>
<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Massachusetts">Massachusetts</a></td>
<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Boston">Boston</a></td> <td>(2020)
675,647</td> <td>(2021 est.) 654,776</td> </tr> <tr> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Michigan">Michigan</a></td> <td><a class="md-crosslink" data-show-preview="true"
href="https://www.britannica.com/place/Lansing-Michigan">Lansing</a></td> <td>(2020) 112,644</td> <td>(2021 est.) 112,684</td> </tr>
<tr> <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Minnesota">Minnesota</a></td> <td>
```

Let's first get just the title of the table.

Note: `<th>` tag defines a header cell in an HTML table

```
titles = soup.find_all('th')
titles
```

```
[<th>state</th>,
 <th>capital</th>,
 <th>population of capital: census</th>,
 <th>population of capital: estimated</th>]
```

Since we do not need the tags, let's clean up the data.

```
titles_list = [title.text for title in titles]
titles_list
```

```
['state',
 'capital',
 'population of capital: census',
 'population of capital: estimated']
```

If the output still contains newline and other symbols that are not needed, you can further clean the data using, for example, `.strip()`

Next, create a dataframe

```
import pandas as pd

df = pd.DataFrame(columns = titles_list)
df
```

state	capital	population of capital: census	population of capital: estimated

Let's scrape the remaining data and fill this table!

```
rows = soup.find_all('tr')
len(rows)
```

```
51
```

The data of our interest are within the scope of **td** tags.

Note: `<td>` tag defines a standard data cell in an HTML table.

```
# -- Long version --
# row_data = []
# for row in rows:
#   row_data.append(row.find_all('td'))

# -- Short version --
row_data = [row.find_all('td') for row in rows]
row_data
```

```
[[],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Alabama-state">Alabama</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Montgomery-Alabama">Montgomery</a>
</td>,
  <td>(2020) 200,603</td>,
  <td>(2021 est.) 198,665</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Alaska">Alaska</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Juneau">Juneau</a></td>,
  <td>(2020) 32,255</td>,
  <td>(2021 est.) 31,973</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Arizona-state">Arizona</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Phoenix-Arizona">Phoenix</a></td>,
  <td>(2020) 1,608,139</td>,
  <td>(2021 est.) 1,624,569</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Arkansas-state">Arkansas</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Little-Rock">Little Rock</a></td>,
  <td>(2020) 202,591</td>,
  <td>(2021 est.) 201,998</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/California-state">California</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Sacramento-California">Sacramento</a>
</td>,
  <td>(2020) 524,943</td>,
  <td>(2021 est.) 525,041</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Colorado-state">Colorado</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Denver">Denver</a></td>,
  <td>(2020) 715,522</td>,
  <td>(2021 est.) 711,463</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Connecticut">Connecticut</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Hartford-Connecticut">Hartford</a>
</td>,
  <td>(2020) 121,054</td>,
  <td>(2021 est.) 120,576</td>],
```

```
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Delaware-state">Delaware</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Dover-Delaware">Dover</a></td>,
  <td>(2020) 39,403</td>,
  <td>(2021 est.) 38,992</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Florida">Florida</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Tallahassee">Tallahassee</a></td>,
  <td>(2020) 196,068</td>,
  <td>(2021 est.) 197,102</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Georgia-state">Georgia</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Atlanta-Georgia">Atlanta</a></td>,
  <td>(2020) 498,715</td>,
  <td>(2021 est.) 496,461</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Hawaii-state">Hawaii</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Honolulu">Honolulu</a></td>,
  <td>(2020) 350,964</td>,
  <td>(2021 est.) 345,510</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Idaho">Idaho</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Boise-Idaho">Boise</a></td>,
  <td>(2020) 235,684</td>,
  <td>(2021 est.) 237,446</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Illinois-state">Illinois</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Springfield-Illinois">Springfield</a>
</td>,
  <td>(2020) 114,394</td>,
  <td>(2021 est.) 113,394</td>],
```

The first row collected has no value, thus an empty list.

```
row_data[0]
```

```
[]
```

Turns out the first row is actually here

```
row_data[1]
```

```
[<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Alabama-state">Alabama</a></td>,
 <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Montgomery-Alabama">Montgomery</a></td>,
 <td>(2020) 200,603</td>,
 <td>(2021 est.) 198,665</td>]
```

However, we only want the text portion.

```
print(row_data[1][0].text)
print(row_data[1][1].text)
print(row_data[1][2].text)
print(row_data[1][3].text)
```

```
Alabama
Montgomery
(2020) 200,603
(2021 est.) 198,665
```

Therefore, more cleaning is necessary.

Add the remaining rows to the dataframe.

But does this code work?

```
row_data
```

```
[[],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Alabama-state">Alabama</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Montgomery-Alabama">Montgomery</a>
</td>,
  <td>(2020) 200,603</td>,
  <td>(2021 est.) 198,665</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Alaska">Alaska</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Juneau">Juneau</a></td>,
  <td>(2020) 32,255</td>,
  <td>(2021 est.) 31,973</td>],
 [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Arizona-state">Arizona</a></td>,
  <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Phoenix-Arizona">Phoenix</a></td>,
  <td>(2020) 1,608,139</td>,
```

```
    <td>(2021 est.) 1,624,569</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Arkansas-state">Arkansas</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Little-Rock">Little Rock</a></td>,
    <td>(2020) 202,591</td>,
    <td>(2021 est.) 201,998</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/California-state">California</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Sacramento-California">Sacramento</a>
  </td>,
    <td>(2020) 524,943</td>,
    <td>(2021 est.) 525,041</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Colorado-state">Colorado</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Denver">Denver</a></td>,
    <td>(2020) 715,522</td>,
    <td>(2021 est.) 711,463</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Connecticut">Connecticut</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Hartford-Connecticut">Hartford</a>
  </td>,
    <td>(2020) 121,054</td>,
    <td>(2021 est.) 120,576</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Delaware-state">Delaware</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Dover-Delaware">Dover</a></td>,
    <td>(2020) 39,403</td>,
    <td>(2021 est.) 38,992</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Florida">Florida</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Tallahassee">Tallahassee</a></td>,
    <td>(2020) 196,068</td>,
    <td>(2021 est.) 197,102</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Georgia-state">Georgia</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Atlanta-Georgia">Atlanta</a></td>,
    <td>(2020) 498,715</td>,
    <td>(2021 est.) 496,461</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Hawaii-state">Hawaii</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Honolulu">Honolulu</a></td>,
    <td>(2020) 350,964</td>,
    <td>(2021 est.) 345,510</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Idaho">Idaho</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Boise-Idaho">Boise</a></td>,
    <td>(2020) 235,684</td>,
    <td>(2021 est.) 237,446</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Illinois-state">Illinois</a></td>,
    <td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Springfield-Illinois">Springfield</a>
  </td>,
    <td>(2020) 114,394</td>,
    <td>(2021 est.) 113,394</td>],
   [<td><a class="md-crosslink" data-show-preview="true" href="https://www.britannica.com/place/Indiana-state">Indiana</a></td>,
```

```
for sentence in row_data[1:]:
    info=[]
    for elem in sentence:
        info.append(elem.text)

    df.loc[len(df)] = info
```

```
df.drop(0, inplace=True)
df
```

	state	capital	population of capital: census	population of capital: estimated
1	Alaska	Juneau	(2020) 32,255	(2021 est.) 31,973
2	Arizona	Phoenix	(2020) 1,608,139	(2021 est.) 1,624,569
3	Arkansas	Little Rock	(2020) 202,591	(2021 est.) 201,998
4	California	Sacramento	(2020) 524,943	(2021 est.) 525,041
5	Colorado	Denver	(2020) 715,522	(2021 est.) 711,463
6	Connecticut	Hartford	(2020) 121,054	(2021 est.) 120,576
7	Delaware	Dover	(2020) 39,403	(2021 est.) 38,992
8	Florida	Tallahassee	(2020) 196,068	(2021 est.) 197,102
9	Georgia	Atlanta	(2020) 498,715	(2021 est.) 496,461
10	Hawaii	Honolulu	(2020) 350,964	(2021 est.) 345,510
11	Idaho	Boise	(2020) 235,684	(2021 est.) 237,446
12	Illinois	Springfield	(2020) 114,394	(2021 est.) 113,394
13	Indiana	Indianapolis	(2020) 887,642	(2021 est.) 882,039
14	Iowa	Des Moines	(2020) 214,133	(2021 est.) 212,031
15	Kansas	Topeka	(2020) 126,587	(2021 est.) 125,963
16	Kentucky	Frankfort	(2020) 28,602	(2021 est.) 28,595
17	Louisiana	Baton Rouge	(2020) 227,470	(2021 est.) 222,185
18	Maine	Augusta	(2020) 18,899	(2021 est.) 18,968
19	Maryland	Annapolis	(2020) 40,812	(2021 est.) 40,687
20	Massachusetts	Boston	(2020) 675,647	(2021 est.) 654,776
21	Michigan	Lansing	(2020) 112,644	(2021 est.) 112,684
22	Minnesota	Saint Paul	(2020) 311,527	(2021 est.) 307,193
23	Mississippi	Jackson	(2020) 153,701	(2021 est.) 149,761
24	Missouri	Jefferson City	(2020) 43,228	(2021 est.) 42,772
25	Montana	Helena	(2020) 32,091	(2021 est.) 33,120
26	Nebraska	Lincoln	(2020) 291,082	(2021 est.) 292,657
27	Nevada	Carson City	(2020) 58,639	(2021 est.) 58,993
28	New Hampshire	Concord	(2020) 43,976	(2021 est.) 44,006
29	New Jersey	Trenton	(2020) 90,871	(2021 est.) 90,457
30	New Mexico	Santa Fe	(2020) 87,505	(2021 est.) 88,193
31	New York	Albany	(2020) 99,224	(2021 est.) 98,617
32	North Carolina	Raleigh	(2020) 467,665	(2021 est.) 469,124
33	North Dakota	Bismarck	(2020) 73,622	(2021 est.) 74,138
34	Ohio	Columbus	(2020) 905,748	(2021 est.) 906,528
35	Oklahoma	Oklahoma City	(2020) 681,054	(2021 est.) 687,725
36	Oregon	Salem	(2020) 175,535	(2021 est.) 177,723
37	Pennsylvania	Harrisburg	(2020) 50,099	(2021 est.) 50,135
38	Rhode Island	Providence	(2020) 190,934	(2021 est.) 189,692
39	South Carolina	Columbia	(2020) 136,632	(2021 est.) 137,541
40	South Dakota	Pierre	(2020) 14,091	(2021 est.) 14,000
41	Tennessee	Nashville	(2020) 689,447	(2021 est.) 678,851
42	Texas	Austin	(2020) 961,855	(2021 est.) 964,177
43	Utah	Salt Lake City	(2020) 199,723	(2021 est.) 200,478
44	Vermont	Montpelier	(2020) 8,074	(2021 est.) 8,002
45	Virginia	Richmond	(2020) 226,610	(2021 est.) 226,604
46	Washington	Olympia	(2020) 55,605	(2021 est.) 55,919

47	West Virginia	Charleston	(2020) 48,864	(2021 est.) 48,018
48	Wisconsin	Madison	(2020) 269,840	(2021 est.) 269,196
49	Wyoming	Cheyenne	(2020) 65,132	(2021 est.) 65,051

TODO:

- Scrape other table from wikipedia
- Generate a new table/tables using dataframe
- Feel free to use other html tags
- Clean & preprocess

Other websites (for instance)

- https://www.timesjobs.com/
- https://www.tripadvisor.com/

```
url = 'https://en.wikipedia.org/wiki/List_of_Bandai_Namco_video_games'
page = requests.get(url)
soup = BeautifulSoup(page.text, 'html')
soup
```

```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vector-
feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-feature-main-
menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled vector-feature-custom-
font-size-clientpref-1 vector-feature-appearance-enabled vector-feature-appearance-pinned-clientpref-1 vector-feature-night-mode-
enabled skin-theme-clientpref-day vector-toc-available" dir="ltr" lang="en">
<head>
<meta charset="utf-8"/>
<title>List of Bandai Namco video games - Wikipedia</title>
<script>(function(){var className="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-
disabled vector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1
vector-feature-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled
vector-feature-custom-font-size-clientpref-1 vector-feature-appearance-enabled vector-feature-appearance-pinned-clientpref-1 vector-
feature-night-mode-enabled skin-theme-clientpref-day vector-toc-available";var cookie=document.cookie.match(/(?:^|;
)enwikimwclientpreferences=([^;]+)/);if(cookie){cookie[1].split('%2C').forEach(function(pref){className=className.replace(new
RegExp('(^| )'+pref.replace(/-clientpref-\w+$|[^\w-]+/g,'')+'-clientpref-\\w+(
|$)'),'$1'+pref+'$2');});}document.documentElement.className=className;}());RLCONF=
{"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":[
"",""],"wgDefaultDateFormat":"dmy","wgMonthNames":
["","January","February","March","April","May","June","July","August","September","October","November","December"],"wgRequestId":"ebe
af0b-4fac-9b97-
f90a79538ea4","wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"List_of_Bandai_Namco_v
of Bandai Namco video
games","wgCurRevisionId":1242499616,"wgRevisionId":1242499616,"wgArticleId":23916522,"wgIsArticle":true,"wgIsRedirect":false,"wgActio
["*"],"wgCategories":["Articles with short description","Short description is different from Wikidata","Bandai Namco games","Video
game lists by
company"],"wgPageViewLanguage":"en","wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgRelevantPageName":"List_of_Bandai
[],
"wgRestrictionMove":[],"wgNoticeProject":"wikipedia","wgCiteReferencePreviewsActive":false,"wgFlaggedRevsParams":{"tags":{"status":
{"levels":1}}},"wgMediaViewerOnClick":true,"wgMediaViewerEnabledByDefault":true,"wgPopupsFlags":6,"wgVisualEditor":
{"pageLanguageCode":"en","pageLanguageDir":"ltr","pageVariantFallbacks":"en"},"wgMFDisplayWikibaseDescriptions":
{"search":true,"watchlist":true,"tagline":false,"nearby":true},"wgWMESchemaEditAttemptStepOversample":false,"wgWMEPageLength":60000,"
[],"wgCentralAuthMobileDomain":false,"wgEditSubmitButtonLabelPublish":true,"wgULSPosition":"interlanguage","wgULSisCompactLinksEnable
["architecture","bitness","brands","fullVersionList","mobile","model","platform","platformVersion"],"GEHomepageSuggestedEditsEnableTo
"wgGETopicsMatchModeEnabled":false,"wgGEStructuredTaskRejectionReasonTextInputEnabled":false,"wgGELevelingUpEnabledForUser":false};RL
{"ext.globalCssJs.user.styles":"ready","site.styles":"ready","user.styles":"ready","ext.globalCssJs.user":"ready","user":"ready","use
["ext.cite.ux-
enhancements","site","mediawiki.page.ready","jquery.tablesorter","jquery.makeCollapsible","mediawiki.toc","skins.vector.js","ext.cent
"ext.centralauth.centralautologin","mmv.head","mmv.bootstrap.autostart","ext.popups","ext.visualEditor.desktopArticleTarget.init","ex
2022","ext.checkUser.clientHints","ext.growthExperiments.SuggestedEditSession","wikibase.sidebar.tracking"];</script>
<script>(RLQ=window.RLQ||[]).push(function(){mw.loader.impl(function(){return["user.options@12s5i",function($,jQuery,require,module)
{mw.user.tokens.set({"patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});
}];});});</script>
<link href="/w/load.php?
lang=en&amp;modules=ext.cite.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimediaBadges%7C
2022" rel="stylesheet"/>
<script async="" src="/w/load.php?lang=en&amp;modules=startup&amp;only=scripts&amp;raw=1&amp;skin=vector-2022"></script>
<meta content="" name="ResourceLoaderDynamicStyles"/>
<link href="/w/load.php?lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector-2022" rel="stylesheet"/>
<meta content="MediaWiki 1.43.0-wmf.19" name="generator"/>
<meta content="origin" name="referrer"/>
<meta content="origin-when-cross-origin" name="referrer"/>
<meta content="max-image-preview:standard" name="robots"/>
<meta content="telephone=no" name="format-detection"/>
<meta content="width=1120" name="viewport"/>
```

```
<meta content="List of Bandai Namco video games - Wikipedia" property="og:title"/>
<meta content="website" property="og:type"/>
```

```
table = soup.find_all('th')

all_title = [title.text.strip('\n') for title in table]
all_title
```

```
['Year',
 'Title',
 'Developer(s)',
 'Platforms',
 'Ref(s)',
 'Title',
 'Year',
 'Platforms',
 'Ref(s)',
 'vteBandai Namco Holdings',
 'Entertainment Unit',
 'Digital Business',
 'Toys & Hobby',
 'IP Production Unit',
 'Amusement Unit',
 'Affiliated Companies',
 'Former subsidiaries',
 'Key people',
 'Defunct',
 'Related',
 'vteVideo game franchises owned by Bandai Namco',
 'Original',
 'Licensed',
 'vteBandai Namco Holdings hardware',
 'Bandai',
 'Namco']
```

```
# I use slicing because this wikipedia has 2 tables which is written by same html tags format: 'Video games' and 'Mobile
# I will focus on the 'Video games' table.
titles = all_title[:5]
titles
```

```
['Year', 'Title', 'Developer(s)', 'Platforms', 'Ref(s)']
```

```
rows = soup.find_all('tr')
rows_data = [row.find_all('td') for row in rows]

game_info = [result for result in rows_data if len(result) == 5]
game_info
```

```
[[<td>2006
  </td>,
  <td><i><a href="/wiki/Kidou_Senshi_Gundam_Seed_Destiny:_Rengou_vs._Z.A.F.T._II" title="Kidou Senshi Gundam Seed Destiny: Rengou
vs. Z.A.F.T. II">Kidou Senshi Gundam Seed Destiny: Rengou vs. Z.A.F.T. II</a></i>
  </td>,
  <td><a href="/wiki/Capcom" title="Capcom">Capcom</a>
  </td>,
  <td><a href="/wiki/PlayStation_2" title="PlayStation 2">PlayStation 2</a>, <a href="/wiki/Arcade_game" title="Arcade
game">Arcade</a>
  </td>,
  <td>
  </td>],
 [<td>2006
  </td>,
  <td><i><a href="/wiki/MotoGP_(2006_video_game)" title="MotoGP (2006 video game)">MotoGP</a></i>
  </td>,
  <td><a href="/wiki/Namco" title="Namco">Namco</a>
  </td>,
  <td><a href="/wiki/PlayStation_Portable" title="PlayStation Portable">PlayStation Portable</a>
  </td>,
  <td>
  </td>],
 [<td>2006
  </td>,
  <td><i><a href="/wiki/Ace_Combat_Zero:_The_Belkan_War" title="Ace Combat Zero: The Belkan War">Ace Combat Zero: The Belkan War</a>
</i>
  </td>,
  <td><a href="/wiki/Namco" title="Namco">Namco</a>
  </td>,
  <td><a href="/wiki/PlayStation_2" title="PlayStation 2">PlayStation 2</a>
  </td>,
```

```
 <td>
 </td>],
[<td>2006
 </td>,
 <td><i><a href="/wiki/Naruto:_Ultimate_Ninja" title="Naruto: Ultimate Ninja">Naruto: Ultimate Ninja</a></i>
 </td>,
 <td><a href="/wiki/CyberConnect2" title="CyberConnect2">CyberConnect2</a>
 </td>,
 <td><a href="/wiki/PlayStation_2" title="PlayStation 2">PlayStation 2</a>
 </td>,
 <td>
 </td>],
[<td>2006
 </td>,
 <td><i><a href="/wiki/Battle_Stadium_D.O.N" title="Battle Stadium D.O.N">Battle Stadium D.O.N</a></i>
 </td>,
 <td><a href="/wiki/Eighting" title="Eighting">Eighting</a>
 <p><a href="/wiki/Q_Entertainment" title="Q Entertainment">Q Entertainment</a>
 </p>
 </td>,
 <td><a href="/wiki/GameCube" title="GameCube">GameCube</a>, <a href="/wiki/PlayStation_2" title="PlayStation 2">PlayStation 2</a>
 </td>,
 <td>
 </td>],
[<td>2006
 </td>,
 <td><i><a href="/wiki/Pac-Man World Rally" title="Pac-Man World Rally">Pac-Man World Rally</a></i>
```

```
bandai = pd.DataFrame(columns=titles)

ind = 0
for record in game_info:
    temp = []
    for elem in record:
        temp.append(elem.text.strip('\n'))

    bandai.loc[ind] = temp
    ind += 1


bandai
```

	Year	Title	Developer(s)	Platforms	Ref(s)
0	2006	Kidou Senshi Gundam Seed Destiny: Rengou vs. Z...	Capcom	PlayStation 2, Arcade	
1	2006	MotoGP	Namco	PlayStation Portable	
2	2006	Ace Combat Zero: The Belkan War	Namco	PlayStation 2	
3	2006	Naruto: Ultimate Ninja	CyberConnect2	PlayStation 2	
4	2006	Battle Stadium D.O.N	Eighting\nQ Entertainment	GameCube, PlayStation 2	
...	...	...	...	...	...
357	2024	Dragon Ball: Sparking! Zero	Spike Chunsoft	Windows, PlayStation 5, Xbox Series X/S	[20]
358	2024	Unknown 9: Awakening	Reflector Entertainment	Windows, PlayStation 5, Xbox Series X/S	
359	2025	Little Nightmares III	Supermassive Games	Windows, Nintendo Switch, PlayStation 4, PlayS...	
360	2025	Fate/EXTRA Record	Type-Moon, Type- Moon studio BB	Windows, Nintendo Switch, PlayStation 4, PlayS...	
361	TBA	Bleach: Rebirth of Souls	Tamsoft	Windows, PlayStation 4, PlayStation 5, Xbox Se...	[21]

362 rows × 5 columns

Clean & Preprocess

```
bandai.drop(columns='Ref(s)', inplace=True)
bandai.head()
```

	Year	Title	Developer(s)	Platforms
0	2006	Kidou Senshi Gundam Seed Destiny: Rengou vs. Z...	Capcom	PlayStation 2, Arcade
1	2006	MotoGP	Namco	PlayStation Portable
2	2006	Ace Combat Zero: The Belkan War	Namco	PlayStation 2
3	2006	Naruto: Ultimate Ninja	CyberConnect2	PlayStation 2
4	2006	Battle Stadium D.O.N	Eighting\nQ Entertainment	GameCube, PlayStation 2

```
# word tokenization
test_df = bandai.copy()

ind = 0
for platform in test_df['Platforms']:
    test_df.loc[ind, 'Platforms'] = platform.split(', ')
    ind += 1

test_df
```

	Year	Title	Developer(s)	Platforms
0	2006	Kidou Senshi Gundam Seed Destiny: Rengou vs. Z...	Capcom	[PlayStation 2, Arcade]
1	2006	MotoGP	Namco	[PlayStation Portable]
2	2006	Ace Combat Zero: The Belkan War	Namco	[PlayStation 2]
3	2006	Naruto: Ultimate Ninja	CyberConnect2	[PlayStation 2]
4	2006	Battle Stadium D.O.N	Eighting\nQ Entertainment	[GameCube, PlayStation 2]
...	...	...	...	...
357	2024	Dragon Ball: Sparking! Zero	Spike Chunsoft	[Windows, PlayStation 5, Xbox Series X/S]
358	2024	Unknown 9: Awakening	Reflector Entertainment	[Windows, PlayStation 5, Xbox Series X/S]
359	2025	Little Nightmares III	Supermassive Games	[Windows, Nintendo Switch, PlayStation 4, Play...
360	2025	Fate/EXTRA Record	Type-Moon, Type- Moon studio BB	[Windows, Nintendo Switch, PlayStation 4, Play...
361	TBA	Bleach: Rebirth of Souls	Tamsoft	[Windows, PlayStation 4, PlayStation 5, Xbox S...

362 rows × 4 columns

I think this dataset needs to preprocess by text tokenization only because it didn't contain the sentence or verb in any columns. I will do the text visualization for extra because I haven't do any much at this point. :)

Text Visualization

```
from itertools import chain

platforms = list(test_df['Platforms'].values)

platforms_list = []
for list_elem in platforms:
    platforms_list = list(chain(platforms_list, list_elem))

platforms_list[:5]
```

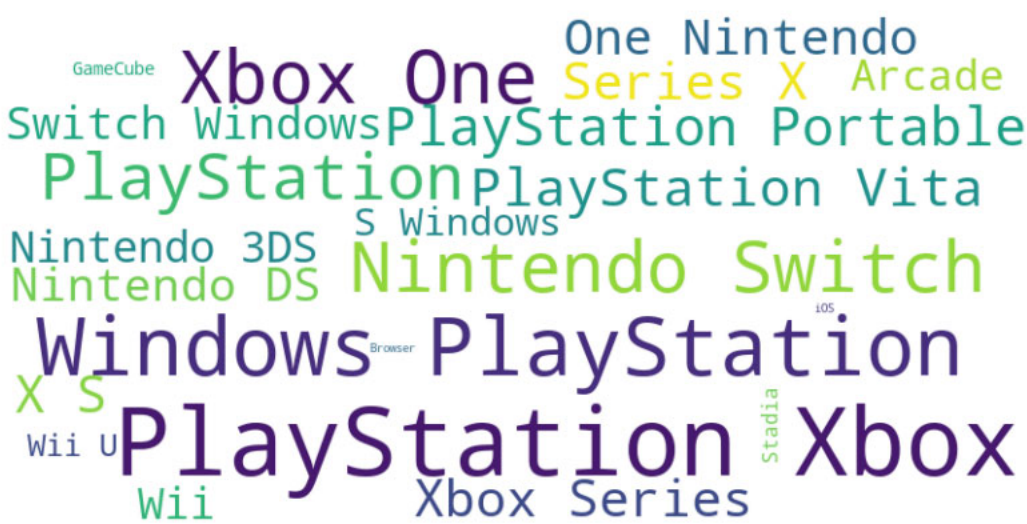
```
['PlayStation 2',
 'Arcade',
 'PlayStation Portable',
 'PlayStation 2',
 'PlayStation 2']
```

```
from nltk.probability import FreqDist

word_freq = FreqDist(platforms_list)

# Convert word frequencies to a DataFrame for seaborn
data = {'Word': list(word_freq.keys()), 'Frequency': list(word_freq.values())}
```

```
df_word_freq = pd.DataFrame(data)

# Sort DataFrame by frequency in descending order
df_word_freq = df_word_freq.sort_values(by='Frequency', ascending=False)

df_word_freq.head(10)
```

	Word	Frequency
12	PlayStation 4	106
4	Windows	104
8	PlayStation 3	82
14	Xbox One	68
13	Nintendo Switch	60
9	Xbox 360	47
2	PlayStation Portable	36
16	PlayStation Vita	35
15	Xbox Series X/S	33
19	PlayStation 5	33

```
import matplotlib.pyplot as plt
from wordcloud import WordCloud

text = ' '.join(platforms_list)
# Generate word cloud
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)

# Display the generated word cloud using matplotlib
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```




```
import seaborn as sns
import matplotlib.pyplot as plt

# Calculate word frequencies
word_freq = FreqDist(platforms_list)

# Convert word frequencies to a DataFrame for seaborn
data = {'Word': list(word_freq.keys()), 'Frequency': list(word_freq.values())}
df_word_freq = pd.DataFrame(data)

# Sort DataFrame by frequency in descending order
```

```
df_word_freq = df_word_freq.sort_values(by='Frequency', ascending=False)

# Plot a bar chart using seaborn
plt.figure(figsize=(12, 6))
sns.barplot(x='Word', y='Frequency', data=df_word_freq.head(20), palette='viridis', hue='Word')
plt.title('Top 20 Most Frequent Words')
plt.xlabel('Words')
plt.ylabel('Frequency')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.show()
```