Using BeautifulSoup and Pandas for Web Scraping

```
# Importung necessary libraries
from bs4 import BeautifulSoup  # For parsing HTML
import requests  # For making HTTP requests
import pandas as pd  # For data manipulation and analysis
```

```
# Defining the URL of the webpage to scrape
url = 'https://en.wikipedia.org/wiki/List_of_largest_companies_in_Africa_by_revenue'
```

```
# Sending a GET request to the URL
page = requests.get(url)

# Parse the HTML content of the page using Beautiful Soup
soup = BeautifulSoup(page.text, 'html.parser')  # Use 'html.parser' for better compatibility
print(soup.prettify())  # Print the prettified HTML for inspection
```

```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vector-feature-page-tools-pinned
 <head>
  <meta charset="utf-8"/>
  <title>
   List of largest companies in Africa by revenue - Wikipedia
  </title>
  <script>
   (function(){var className="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vector-feature-pa
RLSTATE={"ext.globalCssJs.user.styles":"ready","site.styles":"ready","user.styles":"ready","ext.globalCssJs.user":"ready","user":"ready","user.options"
  </script>
  <script>
   (RLQ=window.RLQ||[]).push(function(){mw.loader.impl(function(){return["user.options@12s5i",function($,jQuery,require,module){mw.user.tokens.set({"pat
}];});});});
  </script>
  <link href="/w/load.php?lang=en&amp;modules=ext.cite.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimediame
  <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 1.45.0-wmf.7" 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 largest companies in Africa by revenue - Wikipedia" property="og:title"/>
  <meta content="website" property="og:type"/>
  <link href="//upload.wikimedia.org" rel="preconnect"/>
  <link href="//en.m.wikipedia.org/wiki/List_of_largest_companies_in_Africa_by_revenue" media="only screen and (max-width: 640px)" rel="alternate"/>
```

```
<link href="/w/index.php?title=List_of_largest_companies_in_Africa_by_revenue&amp;action=edit" rel="alternate" title="Edit this page" type="applicatic
<link href="/static/apple-touch/wikipedia.png" rel="apple-touch-icon"/>
<link href="/static/favicon/wikipedia.ico" rel="icon"/>
<link href="/w/rest.php/v1/search" rel="search" title="Wikipedia (en)" type="application/opensearchdescription+xml"/>
<link href="//en.wikipedia.org/w/api.php?action=rsd" rel="EditURI" type="application/rsd+xml"/>
<link href="https://en.wikipedia.org/wiki/List_of_largest_companies_in_Africa_by_revenue" rel="canonical"/>
<link href="https://creativecommons.org/licenses/by-sa/4.0/deed.en" rel="license"/>
<link href="/w/index.php?title=Special:RecentChanges&amp;feed=atom" rel="alternate" title="Wikipedia Atom feed" type="application/atom+xml"/>
<link href="//meta.wikimedia.org" rel="dns-prefetch">
 <link href="auth.wikimedia.org" rel="dns-prefetch"/>
</link>
</head>
<body class="skin--responsive skin-vector skin-vector-search-vue mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-List_of_l
 <a class="mw-jump-link" href="#bodyContent">
  Jump to content
 </a>
 <div class="vector-header-container">
  <header class="vector-header mw-header no-font-mode-scale">
   <div class="vector-header-start">
    <nav aria-label="Site" class="vector-main-menu-landmark">
     <div class="vector-dropdown vector-main-menu-dropdown vector-button-flush-left vector-button-flush-right" id="vector-main-menu-dropdown" title="Ma
      <input aria-haspopup="true" aria-label="Main menu" class="vector-dropdown-checkbox" data-event-name="ui.dropdown-vector-main-menu-dropdown" id="v
      <label aria-hidden="true" class="vector-dropdown-label cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-button--weight-qui
       <span class="vector-icon mw-ui-icon-menu mw-ui-icon-wikimedia-menu">
       </span>
       <span class="vector-dropdown-label-text">
        Main menu
```

```
# Locating the specific table on the page that contains the desired data
# The second table on the page is the one we want to extract
table = soup.find_all('table')[1]  # Adjust index if necessary

# Alternatively, find the table by its class name
# table = soup.find('table', class_='wikitable sortable')
```

```
# Printing the table object to verify it was found correctly
print(table)
```

```
<table class="wikitable sortable">
<tbody><tr>
<th>Rank</th>
<th>Company</th>
<th>Industry</th>
<th>Revenue<br/>(US$ billions)</th>
<th width="150">Headquarters
</th></tr>
<tr>
<td>1</td>
<td><a href="/wiki/Sonatrach" title="Sonatrach">Sonatrach</a></td>
<td>Oil and gas</td>
<td>77.013</td>
```

```
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
<tr>
<td>2</td>
<td><a href="/wiki/Eskom" title="Eskom">Eskom</a></td>
<td>Electric utility</td>
<td>13.941</td>
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
<tr>
<td>3</td>
<td><a href="/wiki/Sasol" title="Sasol">Sasol</a></td>
<td>Chemistry</td>
<td>12.989</td>
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
<tr>
<td>4</td>
<td><a href="/wiki/MTN_Group" title="MTN Group">MTN Group</a></td>
<td>Telecommunications</td>
<td>12.238</td>
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
<tr>
<td>5</td>
<td><a href="/wiki/Shoprite_Holdings_Ltd" title="Shoprite Holdings Ltd">Shoprite Holdings</a></td>
<td>Retail</td>
<td>10.802</td>
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
<tr>
<td>6</td>
<td><a class="mw-redirect" href="/wiki/Nigeria_National_Petroleum_Corporation" title="Nigeria National Petroleum Corporation">Nigeria National Petroleum
<td>Oil and gas</td>
<td>9.706</td>
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
<tr>
<td>7</td>
<td><a href="/wiki/Steinhoff_International" title="Steinhoff International">Steinhoff International</a></td>
<td>Holding</td>
<td>9.704</td>
<td><span class="flagicon"><span class="mw-image-border" typeof="mw:File"><span><img alt="" class="mw-file-element" data-file-height="600" data-file-wid
</td></tr>
```

```
# Extracting the table headers (titles)
titles = table.find_all('th')
# Create a list of titles for the DataFrame
table_titles = [title.text.strip() for title in titles]
```

```
# Initialize a DataFrame with the extracted titles as columns
df = pd.DataFrame(columns=table_titles)
```

```
# Extracting all rows of the table
column_data = table.find_all('tr')
```

```
# Loop through each row in the table, skipping the header row
for row in column_data[1:]:
    # Find all data cells in the row
    row_data = row.find_all('td')
    # Extract text from each cell and strip whitespace
    individual_row_data = [x.text.strip() for x in row_data]
    # Append the row data to the DataFrame
    df.loc[len(df)] = individual_row_data  # Use len(df) to get the next index

df
```

	Rank	Company	Industry	Revenue(US$ billions)	Headquarters
0	1	Sonatrach	Oil and gas	77.013	Algeria
1	2	Eskom	Electric utility	13.941	South Africa
2	3	Sasol	Chemistry	12.989	South Africa
3	4	MTN Group	Telecommunications	12.238	South Africa
4	5	Shoprite Holdings	Retail	10.802	South Africa
...	...	...	...	...	...
95	96	Blue Label Telecoms	Telecommunications	1.442	South Africa
96	97	Kibali Gold Mine	Mining	1.440	DR Congo
97	98	Aveng	Conglomerate	1.425	South Africa
98	99	Murray and Roberts Holdings	Construction	1.422	South Africa
99	100	Rustenburg	Mining	1.240	South Africa

100 rows × 5 columns

```
# Saving the DataFrame to a CSV file
df.to_csv(r'C:\Users\user\Downloads\Web scraping\Companies_in_Africa.csv', index=False)
```