# Scraping Data From a Real Website + Pandas

In [3]:
```
from bs4 import BeautifulSoup
import requests
```

In [4]:
```
url = 'https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States_by_revenue'
```

In [9]:
```
page = requests.get(url)

soup = BeautifulSoup(page.text, 'html')
```

In [10]:
```
print(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-pinn
ed-disabled vector-feature-toc-pinned-clientpref-1 vector-feature-main-menu-pinned-disabled ve
ctor-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled vector-fe
ature-custom-font-size-clientpref-0 vector-feature-appearance-disabled vector-feature-appearan
ce-pinned-clientpref-0 vector-feature-night-mode-disabled skin-theme-clientpref-day vector-toc
-available" dir="ltr" lang="en">
<head>
<meta charset="utf-8"/>
<title>List of largest companies in the United States 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-sticky-header-disabled vector-fea
ture-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-feature-main-men
u-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-conte
nt-enabled vector-feature-custom-font-size-clientpref-0 vector-feature-appearance-disabled vec
tor-feature-appearance-pinned-clientpref-0 vector-feature-night-mode-disabled skin-theme-clien
tpref-day vector-toc-available";var cookie=document.cookie.match(/(?:^|; )enwikimwclientprefer
ences=([^;]+)/);if(cookie){cookie[1].split('%2C').forEach(function(pref){className=className.r
```

In [11]:
```
soup.find('table', class_='wikitable sortable')
```

Out[11]:
```
<table class="wikitable sortable">
<caption>
</caption>
<tbody><tr>
<th>Rank
</th>
<th>Name
</th>
<th>Industry
</th>
<th>Revenue <br/>(USD millions)
</th>
<th>Revenue growth
</th>
<th>Employees
</th>
<th>Headquarters
</th></tr>
<tr>
```

In [12]:
```
table = soup.find_all('table')[1]
```

In [13]:
```
print(table)
```

```
<table class="wikitable sortable">
<caption>
</caption>
<tbody><tr>
<th>Rank
</th>
<th>Name
</th>
<th>Industry
</th>
<th>Revenue <br/>(USD millions)
</th>
<th>Revenue growth
</th>
<th>Employees
</th>
<th>Headquarters
</th></tr>
<tr>
```

In [14]:
```
world_titles = table.find_all('th')
```

In [15]:
```
world_table_titles = [title.text.strip() for title in world_titles]

print(world_table_titles)
```

```
['Rank', 'Name', 'Industry', 'Revenue (USD millions)', 'Revenue growth', 'Employees', 'Headquart
ers']
```

In [16]:
```
import pandas as pd
```

In [17]:
```
df = pd.DataFrame(columns = world_table_titles)
df
```

Out[17]:

Rank	Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters

In [22]:
```
column_data = table.find_all('tr')
```

In [26]:
```
for row in column_data[1:]:
    row_data = row.find_all('td')
    individual_row_data = [data.text.strip() for data in row_data]

    length = len(df)
    df.loc[length] = individual_row_data
```

In [27]:
```
df
```

Out[27]:

	Rank	Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters
0	1	Walmart	Retail	611,289	6.7%	2,100,000	Bentonville, Arkansas
1	2	Amazon	Retail and cloud computing	513,983	9.4%	1,540,000	Seattle, Washington
2	3	ExxonMobil	Petroleum industry	413,680	44.8%	62,000	Spring, Texas
3	4	Apple	Electronics industry	394,328	7.8%	164,000	Cupertino, California
4	5	UnitedHealth Group	Healthcare	324,162	12.7%	400,000	Minnetonka, Minnesota
...	...	...	...	...	...	...	...
95	96	Best Buy	Retail	46,298	10.6%	71,100	Richfield, Minnesota
96	97	Bristol-Myers Squibb	Pharmaceutical industry	46,159	0.5%	34,300	New York City, New York
97	98	United Airlines	Airline	44,955	82.5%	92,795	Chicago, Illinois
98	99	Thermo Fisher Scientific	Laboratory instruments	44,915	14.5%	130,000	Waltham, Massachusetts
99	100	Qualcomm	Technology	44,200	31.7%	51,000	San Diego, California

100 rows × 7 columns

In [30]:
```
lf.to_csv(r'D:\COURSES\YOUTUBE\ALEX THE ANALYST\PYTHON\Web scraping companies.csv', index = False)
```

In [ ]: