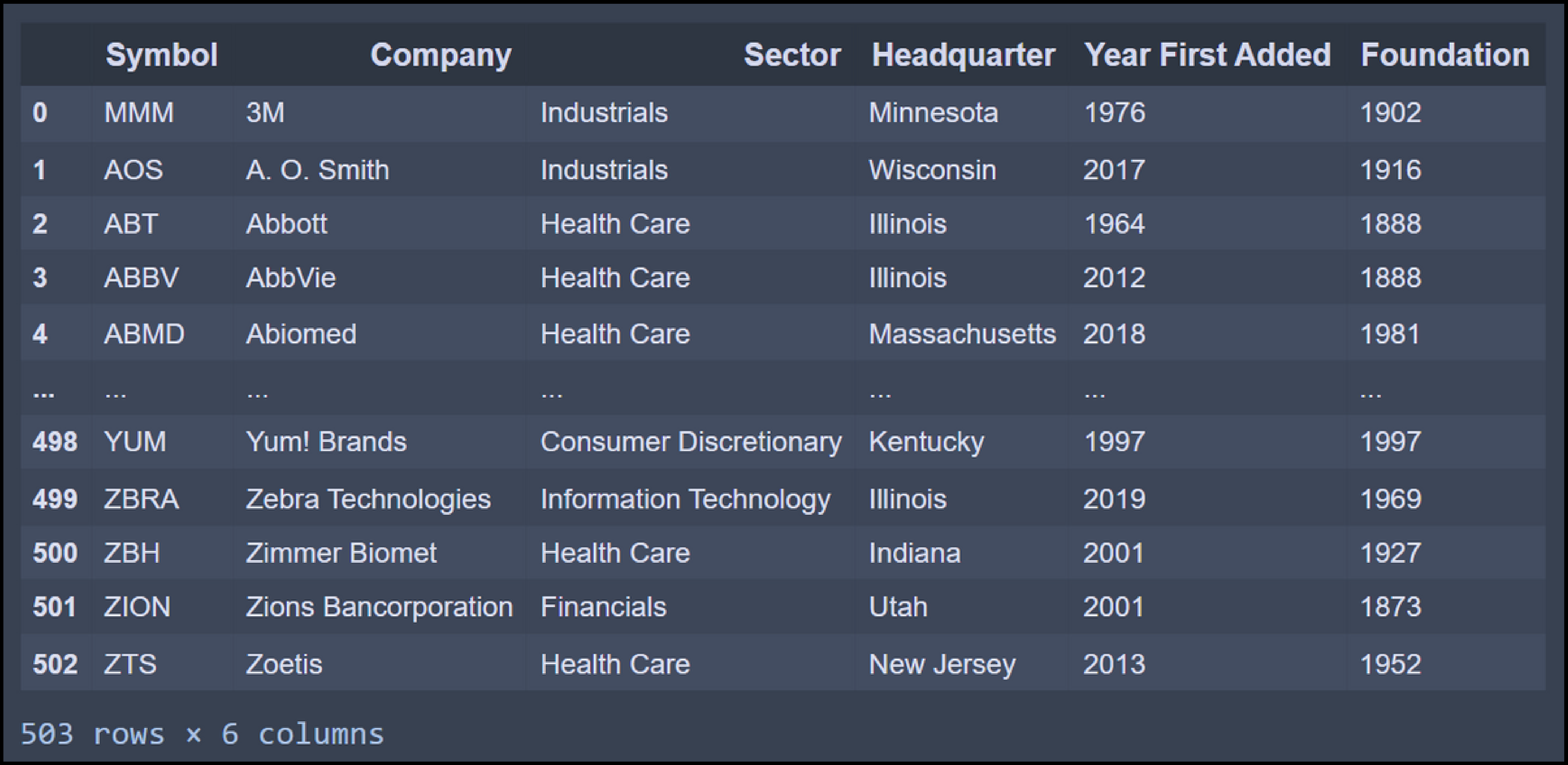
Python

Web scrape S&P 500 index data with BeautifulSoup & Requests

STANDARD
&POOR'S 500

Web scrape features of S&P 500 index companies



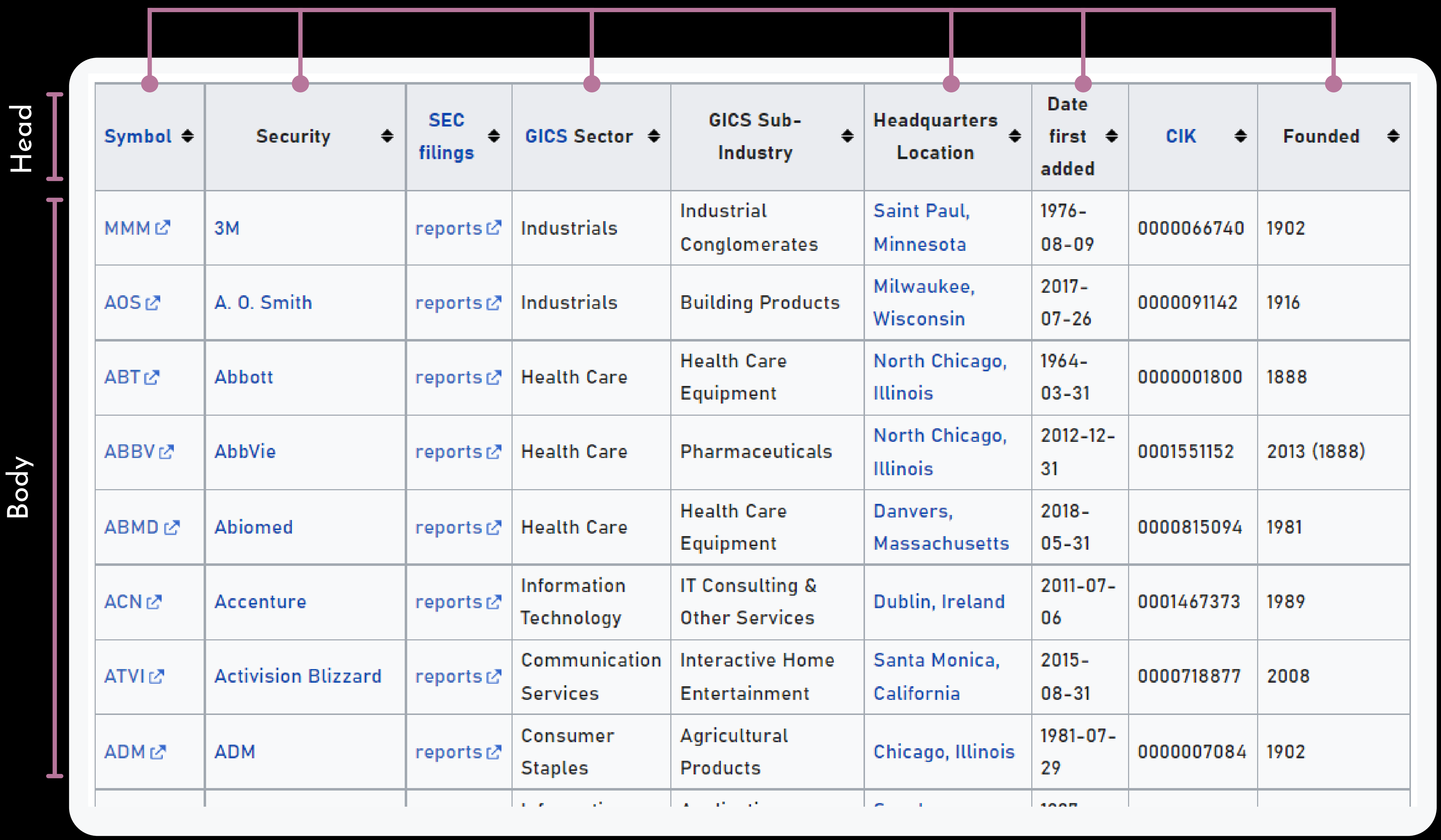
	Symbol	Company	Sector	Headquarter	Year First Added	Foundation
0	MMM	3M	Industrials	Minnesota	1976	1902
1	AOS	A. O. Smith	Industrials	Wisconsin	2017	1916
2	ABT	Abbott	Health Care	Illinois	1964	1888
3	ABBV	AbbVie	Health Care	Illinois	2012	1888
4	ABMD	Abiomed	Health Care	Massachusetts	2018	1981
...	...	...	...	...	...	...
498	YUM	Yum! Brands	Consumer Discretionary	Kentucky	1997	1997
499	ZBRA	Zebra Technologies	Information Technology	Illinois	2019	1969
500	ZBH	Zimmer Biomet	Health Care	Indiana	2001	1927
501	ZION	Zions Bancorporation	Financials	Utah	2001	1873
502	ZTS	Zoetis	Health Care	New Jersey	2013	1952

```
503 rows × 6 columns
```

Step 1:

Identify S&P 500 features of interest

S&P 500 index is a market-capitalization-weighted index of 503 leading publicly traded companies (stocks) in the U.S.. We'll web scrape these 6 (from 9) features for every stock



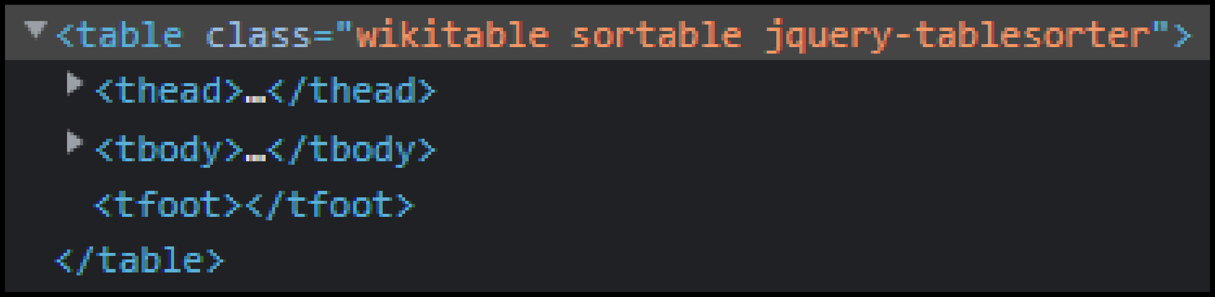
Symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Headquarters Location	Date first added	CIK	Founded
MMM	3M	reports	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1976-08-09	0000066740	1902
AOS	A. O. Smith	reports	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	0000091142	1916
ABT	Abbott	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
ABBV	AbbVie	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ABMD	Abiomed	reports	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	0000815094	1981
ACN	Accenture	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989
ATVI	Activision Blizzard	reports	Communication Services	Interactive Home Entertainment	Santa Monica, California	2015-08-31	0000718877	2008
ADM	ADM	reports	Consumer Staples	Agricultural Products	Chicago, Illinois	1981-07-29	0000007084	1902

Source: Wikipedia "List of S&P 500 companies"

Step 2:

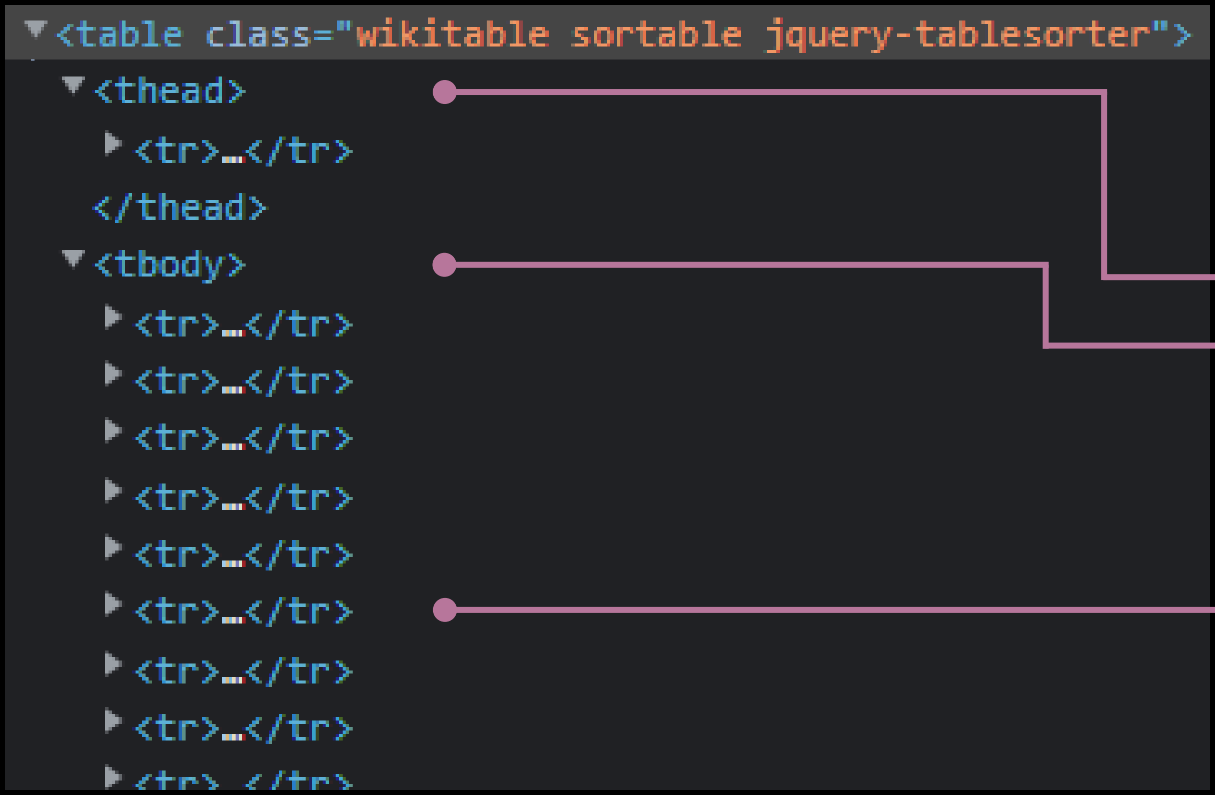
Identify table inside html

We use *requests* library to read data from the page and *BeautifulSoup* library to convert this data to a certain object in order to work with HTML efficiently



```
▼<table class="wikitable sortable jquery-tablesorter">
  ▶<thead>…</thead>
  ▶<tbody>…</tbody>
   <tfoot></tfoot>
 </table>
```

The table is found using 'wikitable sortable' class



```
▼<table class="wikitable sortable jquery-tablesorter">
  ▼<thead>
    ▶<tr>…</tr>
   </thead>
  ▼<tbody>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
```

As we see in the previous step, the table includes:

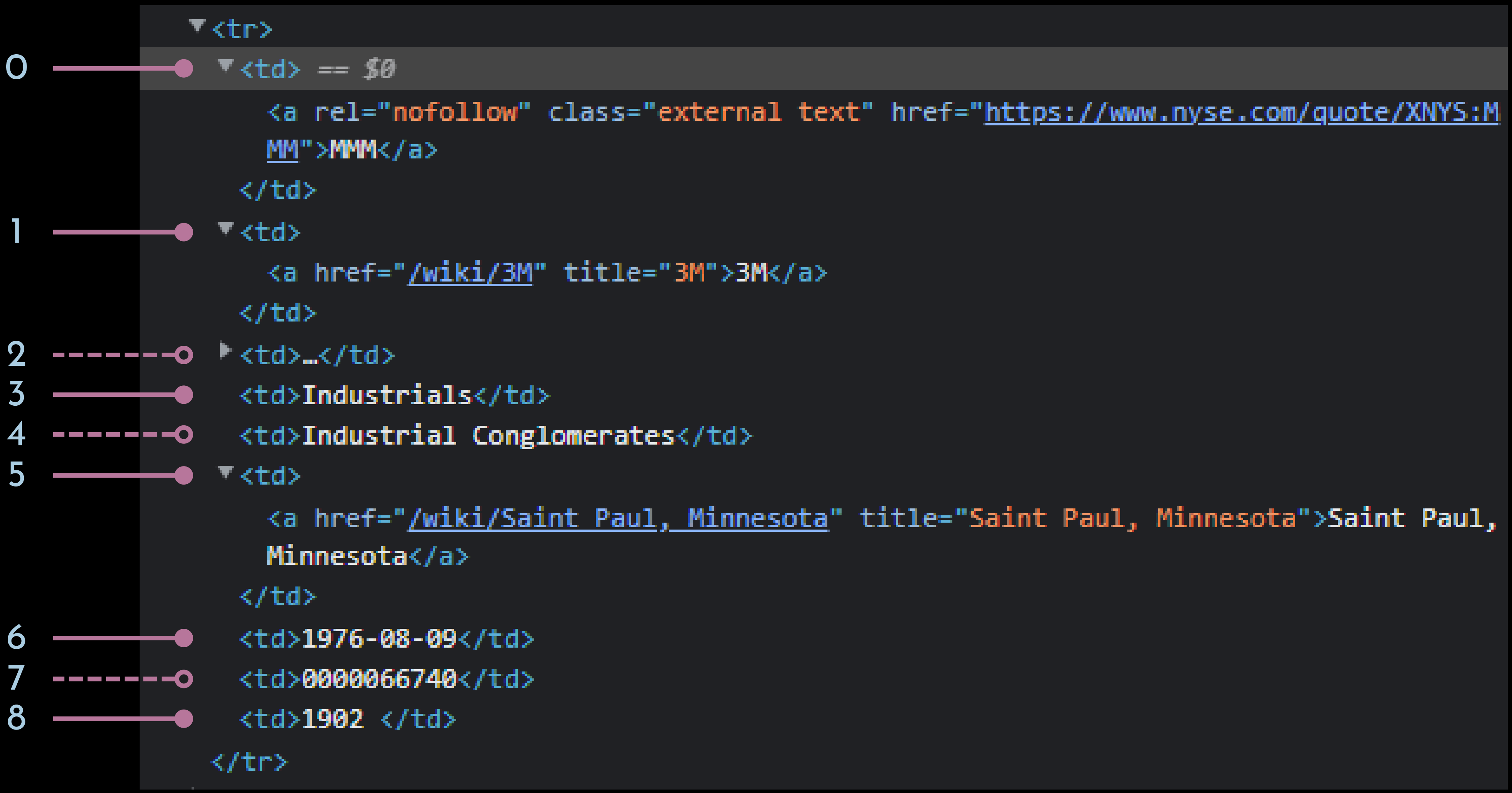
- head: <thead> element
- body: <tbody> element

every <tr> element stands for a row of the table

Step 3:

Understand <tr> element structure

Inside every <tr> element, there are <td> elements that refer to a feature (*white text*), but we are only interested in 6 features (order 0, 1, 3, 5, 6, 8)



```
▼<tr>
  ▼<td> == $0
     <a rel="nofollow" class="external text" href="https://www.nyse.com/quote/XNYS:M
     MM">MMM</a>
   </td>
  ▼<td>
     <a href="/wiki/3M" title="3M">3M</a>
   </td>
  ▶<td>…</td>
   <td>Industrials</td>
   <td>Industrial Conglomerates</td>
  ▼<td>
     <a href="/wiki/Saint_Paul,_Minnesota" title="Saint Paul, Minnesota">Saint Paul,
     Minnesota</a>
   </td>
   <td>1976-08-09</td>
   <td>0000066740</td>
   <td>1902 </td>
 </tr>
```

Step 4:

Iterate over every row or <tr> element

The 6 features are web scraped from each row starting from the second one because the header is not considered

```
for row in tqdm(table.findAll('tr')[1:]):
    ticker = row.findAll('td')[0].text
    company = row.findAll('td')[1].text
    industry = row.findAll('td')[3].text
    headquarter = row.findAll('td')[5].text
    date_1st_added = row.findAll('td')[6].text
    year_founded = row.findAll('td')[8].text
```

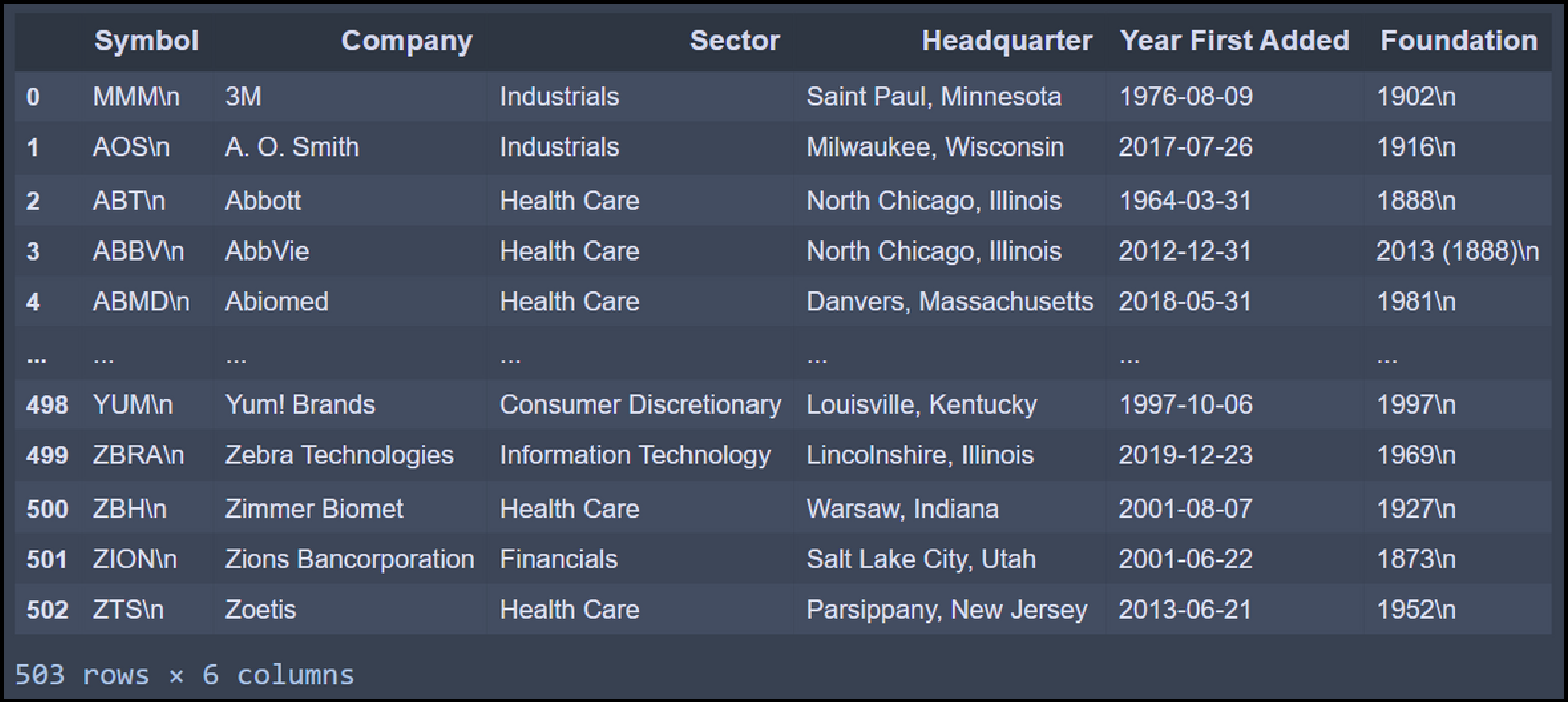
Extracted data is stored in lists to be later converted to a dataframe

```
    tickers.append(ticker)
    companies.append(company)
    industries.append(industry)
    headquarters.append(headquarter)
    dates_1st_added.append(date_1st_added)
    years_founded.append(year_founded)
```

Step 5:

Convert lists to dataframe

Lists containing extracted data are converted to a data frame by specifying the column names. Note that there are 503 stocks



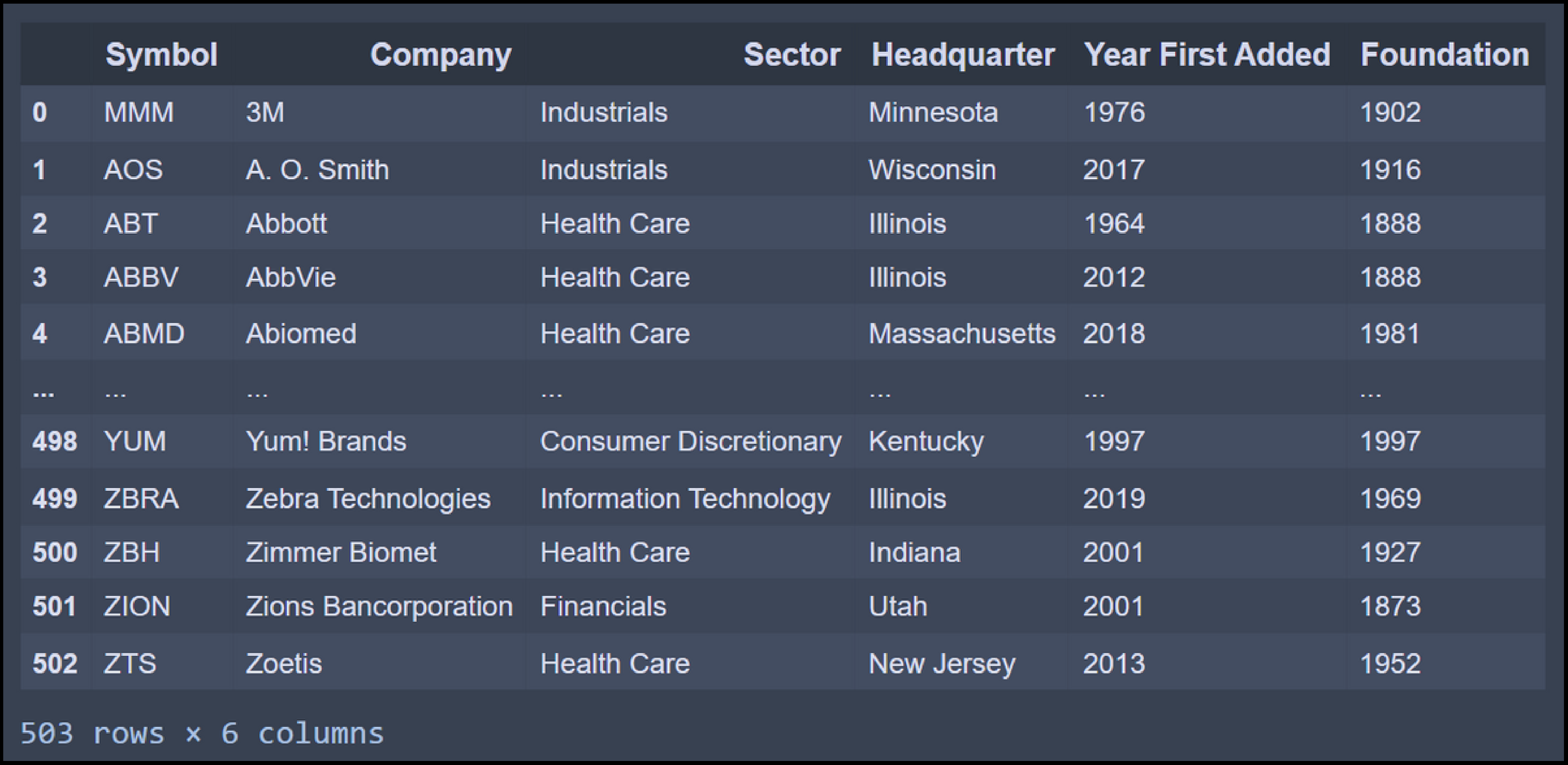
	Symbol	Company	Sector	Headquarter	Year First Added	Foundation
0	MMM\n	3M	Industrials	Saint Paul, Minnesota	1976-08-09	1902\n
1	AOS\n	A. O. Smith	Industrials	Milwaukee, Wisconsin	2017-07-26	1916\n
2	ABT\n	Abbott	Health Care	North Chicago, Illinois	1964-03-31	1888\n
3	ABBV\n	AbbVie	Health Care	North Chicago, Illinois	2012-12-31	2013 (1888)\n
4	ABMD\n	Abiomed	Health Care	Danvers, Massachusetts	2018-05-31	1981\n
...	...	...	...	...	...	...
498	YUM\n	Yum! Brands	Consumer Discretionary	Louisville, Kentucky	1997-10-06	1997\n
499	ZBRA\n	Zebra Technologies	Information Technology	Lincolnshire, Illinois	2019-12-23	1969\n
500	ZBH\n	Zimmer Biomet	Health Care	Warsaw, Indiana	2001-08-07	1927\n
501	ZION\n	Zions Bancorporation	Financials	Salt Lake City, Utah	2001-06-22	1873\n
502	ZTS\n	Zoetis	Health Care	Parsippany, New Jersey	2013-06-21	1952\n

```
503 rows × 6 columns
```

Step 6:

Data cleaning

It's necessary to clean '\n', keep city/country in *Headquarter* column and keep year in *Year First Added* column. Now, we are ready to perform data visualization using this data



	Symbol	Company	Sector	Headquarter	Year First Added	Foundation
0	MMM	3M	Industrials	Minnesota	1976	1902
1	AOS	A. O. Smith	Industrials	Wisconsin	2017	1916
2	ABT	Abbott	Health Care	Illinois	1964	1888
3	ABBV	AbbVie	Health Care	Illinois	2012	1888
4	ABMD	Abiomed	Health Care	Massachusetts	2018	1981
...	...	...	...	...	...	...
498	YUM	Yum! Brands	Consumer Discretionary	Kentucky	1997	1997
499	ZBRA	Zebra Technologies	Information Technology	Illinois	2019	1969
500	ZBH	Zimmer Biomet	Health Care	Indiana	2001	1927
501	ZION	Zions Bancorporation	Financials	Utah	2001	1873
502	ZTS	Zoetis	Health Care	New Jersey	2013	1952

```
503 rows × 6 columns
```

More projects:

linktr.ee/sandreke99

sandreke99