Wikipedia Top Companies Web Scrapping

A project Submitted to the

IT Vedant Institute, Thane

Data Science & Analytics With AI

Python Web Scrapping Project

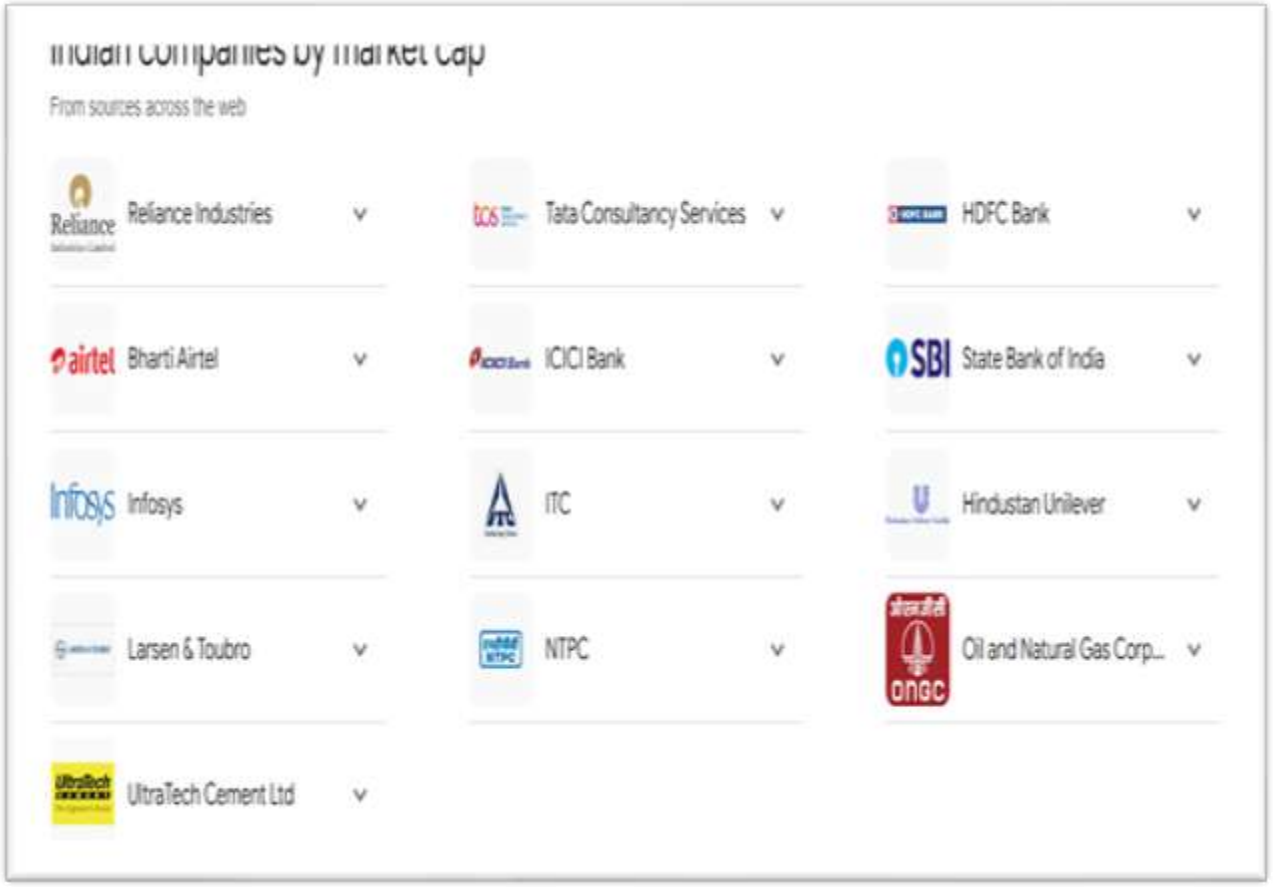
BY

Om Suryawanshi

Under the Guidence of

Mr. Sameer Warsolkar

Top Companies In India



Using Pyhon And Beautifulsoup

Description:

Designed and implemented a Python project using BeautifulSoup, Request to scrape data on top companies in India. Extracted information such as rankings and industry sectors. Processed and analyzed the data to identify key trends. Created a comprehensive PDF report summarizing the findings, incorporating tables and visualizations for clear presentation. This project highlights proficiency in web scraping, data analysis, and report generation.

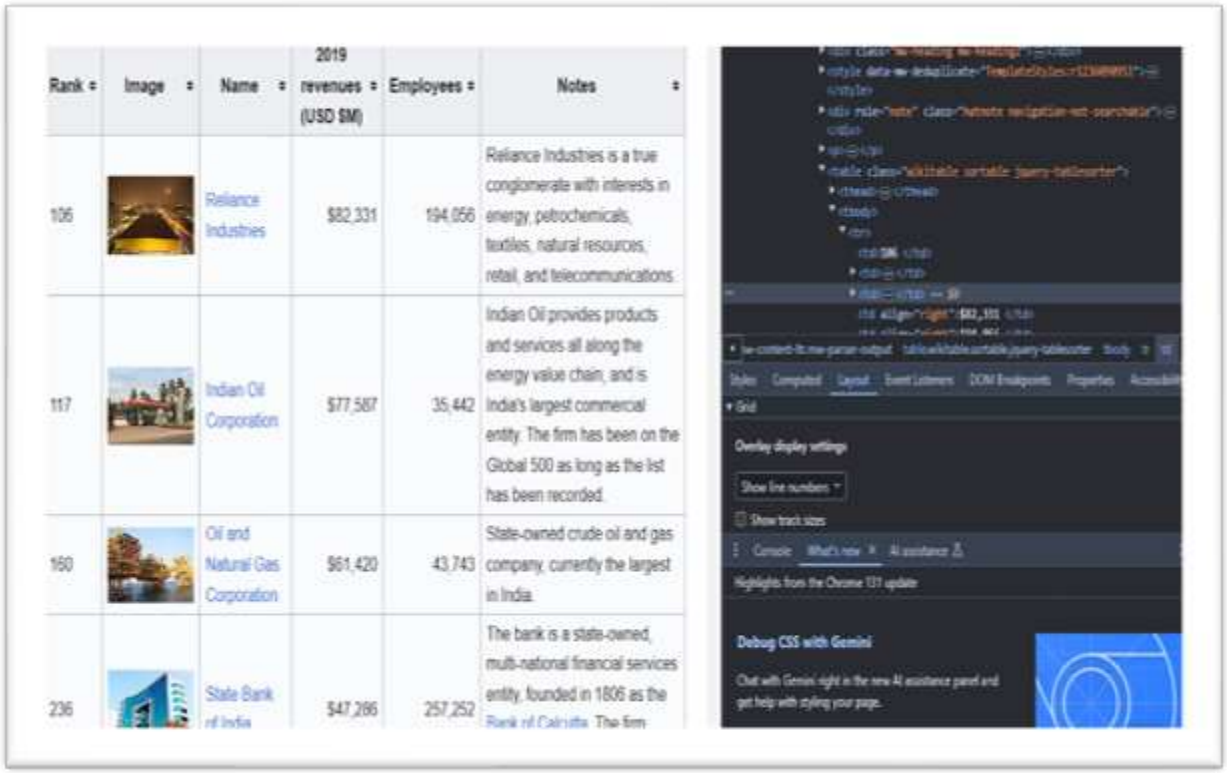
Outline:

From this site, we are going to grab the following information:

- Company
- Industry
- Sector
- Main Office Located
- Start-up Date

Steps:

Choose the Website and Webpage URL
Inspect the Website



Rank	Image	Name	2019 revenues (USD $M)	Employees	Notes
106		Reliance Industries	$82,331	194,056	Reliance Industries is a true conglomerate with interests in energy, petrochemicals, textiles, natural resources, retail, and telecommunications
117		Indian Oil Corporation	$77,587	35,442	Indian Oil provides products and services all along the energy value chain, and is India's largest commercial entity. The firm has been on the Global 500 as long as the list has been recorded.
160		Oil and Natural Gas Corporation	$61,420	43,743	State-owned crude oil and gas company, currently the largest in India.
236		State Bank of India	$47,286	257,252	The bank is a state-owned, multi-national financial services entity, founded in 1806 as the Bank of Calcutta. The firm

Installing the important libraries

Python has several web scrapping libraries, We will use the following libraries

- **Request**
- **BeautifulSoup**
- **Pandas**

Write the python code for exporting the extracted data

Top Companies Web Scrapping Coding

Accessing the Website using request and beautifulSoup libraries,

And using html.parser

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

```
url = "https://en.wikipedia.org/wiki/List_of_companies_of_India"
response = requests.get(url)
response
```

```
<Response [200]>
```

```
[3]: soup = BeautifulSoup(response.content, "html.parser")
```

```
[4]: # Find all td elements
     td_elements = soup.find_all("td")
     # td_elements
```

```
[5]: # Extract text from each td element
     import re
     list_general=[]
     for td in td_elements:

         text = td.get_text()
         text1=re.sub("\n"," ",text)
         list_general.append(text1)

     list_general
```

```
 'Oil and Natural Gas Corporation ',
 '$61,428 ',
 '43,743 ',
 'State-owned crude oil and gas company, currently the largest in India. ',
 '236 ',
 ' ',
 'State Bank of India ',
 '$47,286 ',
 '257,252 ',
 'The bank is a state-owned, multi-national financial services entity, founded in 1806 as the Bank of Calcutta. The firm operates in more than 36 coun
tries. ',
 '275 ',
 ' ',
 'Bharat Petroleum ',
 '$42,935 ',
 '12,865 ',
 'Government controlled oil and gas company in Mumbai, operating large refineries in Kochi and Mumbai. ',
 '63 Moons Technologies ',
```

Let us find their index in the list

```
[6]: # The data we want to extract starts with name: Moons Technologies
     # The last name in the table is : Zydus Lifescience
     search_first_row_text='63 Moons Technologies '
     search_last_row_text='Zydus Lifesciences '

[7]: # Let us find their index in the list
     search_index_first_row=list_general.index(search_first_row_text)
     search_index_last_row=list_general.index(search_last_row_text)

     print("The index of the'63 Moons Technologies' in the above list is: ",search_index_first_row)
     print("The index of the 'Zydus Lifesciences' in the above list is: ",search_index_last_row)

     The index of the'63 Moons Technologies' in the above list is:  30
     The index of the 'Zydus Lifesciences' in the above list is:  3958
```

Extracting Table-Only content from the above list.

```
[8]: # Extracting Table-Only content from the above list
     new_list=[]
     for i in range(search_index_first_row-8,search_index_last_row-5):
         new_list.append(list_general[i])
     print(new_list)

     ['63 Moons Technologies ', 'Technology ', 'Software ', 'Mumbai ', '1988 ', 'Financial technology ', 'P ', 'A ', 'Aban Offshore ', 'Energy ', 'Oil & ga
     s offshore drilling & other services ', 'Chennai ', '1986 ', 'Oil, petrochemical ', 'P ', 'A ', 'ABG Shipyard ', 'Industrials ', 'Shipbuilding ', 'Ahm
     edabad ', '1985 ', 'Ship engineering ', 'P ', 'A ', 'ABP Group ', 'Consumer services ', 'Broadcasting & entertainment ', 'Kolkata ', '1922 ', 'Media,
     news ', 'P ', 'A ', 'ACC ', 'Industrials ', 'Cement ', 'Mumbai ', '1936 ', 'Cement ', 'P ', 'A ', 'ACG Group ', 'Industrials ', 'Machinery: specialty
     ', 'Mumbai ', '1961 ', 'Pharmaceutical industry machinery & medicinal materials manufacturer ', 'P ', 'A ', 'Acko General Insurance ', 'Insurance ',
     'Full line insurance ', 'Mumbai ', '2016 ', 'Insurance ', 'P ', 'A ', 'Action Construction Equipment ', 'Industrials ', 'Machinery: construction & han
     dling ', 'Faridabad ', '1995 ', 'Bulldozers, cranes ', 'P ', 'A ', 'Action Group ', 'Conglomerate ', '— ', 'New Delhi ', '1972 ', 'Apparel, chemicals,
     retail, steel ', 'P ', 'A ', 'Adani Group ', 'Conglomerate ', '— ', 'Ahmedabad ', '1988 ', 'Conglomerate ', 'P ', 'A ', 'Adani Power ', 'Utilities ',
     'Alternative electricity ', 'Mumbai ', '1910 ', 'Thermal & solar power generation, part of Adani Group ', 'P ', 'A ', 'Aditya Birla Fashion and Retail
     ', 'Retail ', 'Apparel retailers ', 'Mumbai ', '1997 ', 'Clothing retailer ', 'P ', 'A ', 'Aditya Birla Group ', 'Conglomerate ', '— ', 'Mumbai ', '18
     57 ', 'Agribusiness, cement, chemicals ', 'P ', 'A ', 'Ador Group ', 'Conglomerate ', '— ', 'Mumbai ', '1977 ', 'Electronics, energy, chemicals ', 'P
     ', 'A ', 'Advanced Weapons and Equipment India ', 'Industrials ', 'Defense ', 'Kanpur ', '2021 ', 'Weapons ', 'S ', 'A ', 'Afcons Infrastructure ', 'I
     ndustrials ', 'Construction ', 'Mumbai ', '1959 ', 'Constructions ', 'P ', 'A ', 'Agriculture Insurance Company of India ', 'Insurance ', 'Full line i
     nsurance ', 'New Delhi ', '2002 ', 'Insurance ', 'S ', 'A ', 'Air Costa ', 'Consumer services ', 'Airlines ', 'Vijayawada ', '2013 ', 'Defunct 2017 ',
     'P ', 'D ', 'Air India ', 'Consumer services ', 'Airlines ', 'New Delhi ', '1932 ', 'Airline, part of Tata Sons ', 'P ', 'A ', 'Air India Express ',
     'Consumer services ', 'Airlines ', 'Kochi ', '2005 ', 'Low-cost airline, part of Air India ', 'P ', 'A ', 'Ajanta Group ', 'Conglomerate ', '— ', 'Ahm
     edabad ', '1971 ', 'Textile, transport, construction machinery ', 'P ', 'A ', 'Alkem Laboratories ', 'Health care ', 'Pharmaceuticals ', 'Mumbai ', '1
     973 ', 'Pharma ', 'P ', 'A ', 'Allcargo Logistics ', 'Industrials ', 'Transportation services ', 'Mumbai ', '1993 ', 'Logistics & transportation ', 'P
```

Now accessing the Companies Name

[5]:
```
# Extract text from each td element
import re
list_general=[]
for td in td_elements:

    text = td.get_text()
    text1=re.sub("\n"," ",text)
    list_general.append(text1)

list_general
```

[5]:
```
['106 ',
 ' ',
 'Reliance Industries ',
 '$82,331 ',
 '194,056 ',
 'Reliance Industries is a true conglomerate with interests in energy, petrochemicals, textiles, natural resources, retail, and telecommunications. ',
 '117 ',
 ' ',
 'Indian Oil Corporation ',
 '$77,587 ',
 '35,442 ',
 "Indian Oil provides products and services all along the energy value chain, and is India's largest commercial entity.  The firm has been on the Glob
al 500 as long as the list has been recorded. ",
 '160 ',
 ' ',
 'Oil and Natural Gas Corporation ',
 '$61,420 ',
 '43 763 '
```

[9]:
```
# Making seporate list of names
Company_Names=[]
for i in range(0,len(new_list),8):
    Company_Names.append(new_list[i])
print(Company_Names)
```

```
['63 Moons Technologies ', 'Aban Offshore ', 'ABG Shipyard ', 'ABP Group ', 'ACC ', 'ACG Group ', 'Acko General Insurance ', 'Action Construction Equipm
ent ', 'Action Group ', 'Adani Group ', 'Adani Power ', 'Aditya Birla Fashion and Retail ', 'Aditya Birla Group ', 'Ador Group ', 'Advanced Weapons and
Equipment India ', 'Afcons Infrastructure ', 'Agriculture Insurance Company of India ', 'Air Costa ', 'Air India ', 'Air India Express ', 'Ajanta Group
', 'Alkem Laboratories ', 'Allcargo Logistics ', 'Amartex ', 'Ambuja Cements ', 'Amrutanjan Healthcare ', 'Amul ', 'Angel One ', 'Apollo Hospitals ', 'A
pollo Tyres ', 'Aptech ', 'Archies ', 'Arise India ', 'Armoured Vehicles Nigam ', 'Artemis Hospital ', 'Arvind ', 'Ashok Leyland ', 'Asia MotorWorks ',
'Asian Paints ', 'Atul ', 'Atul Auto ', 'Avantha Group ', 'Axis Bank ', 'Bajaj Auto ', 'Bajaj Group ', 'Balaji Telefilms ', 'Balaji Wafers ', 'Ballarpur
Industries ', 'Bank of Baroda ', 'Bank of India ', 'Bank of Maharashtra ', 'Bellatrix Aerospace ', 'BEML ', 'BGR Energy Systems ', 'Bharat Aluminium Com
pany ', 'Bharat Biotech ', 'Bharat Dynamics Limited ', 'Bharat Electronics ', 'Bharat Forge ', 'Bharat Heavy Electricals Limited ', 'Bharat Petroleum ',
'Bharat Sanchar Nigam Limited ', 'BHAVINI ', 'Bharti Airtel ', 'Biocon ', 'Blue Star ', 'BluSmart Mobility ', 'Bombay Dyeing ', 'Bonn Group of Industrie
s ', 'Borosil ', 'BPL Group ', 'BrahMos Aerospace ', 'Brigade Enterprises ', 'Britannia Industries ', 'Café Coffee Day ', 'Canara Bank ', 'CaratLane ',
'CEAT ', 'Celkon ', 'Central Bank of India ', 'Centum Electronics ', 'Century Plyboards ', 'CESC Limited ', 'CG Power and Industrial Solutions ', 'Chenn
ai Petroleum Corporation ', 'Cipla ', 'City Union Bank ', 'CMC ', 'Coal India ', 'Cochin Minerals and Rutile Limited ', 'Cochin Shipyard ', 'Computer Ag
e Management Services ', 'Container Corporation of India ', 'Coromandel International ', 'Cosmic Circuits ', 'Crossword Bookstores ', 'CSB Bank ', 'Cyie
nt ', 'Dabur ', 'Damodar Valley Corporation ', 'Datamatics ', 'Deccan Charters ', 'Dedicated Freight Corridor Corporation of India ', 'Dempo ', 'Dena Ba
nk ', 'Dish TV ', 'Dixon Technologies ', 'DLF ', 'DMart ', "Dr. Reddy's Laboratories ", 'Dynamatic Technologies ', 'EaseMyTrip ', 'Edelweiss Group ', 'E
icher Motors ', 'EID Parry ', 'Electronics Corporation of India Limited ', 'eLitmus ', 'Emami ', 'Emcure Pharmaceuticals ', 'Engineers India ', 'English
Indian Clays ', 'Escorts Kubota Limited ', 'Essar Group ', 'Essel Group ', 'Eureka Forbes ', 'Eveready Industries India ', 'Exide Industries ', 'Exide L
ife Insurance ', 'Federal Bank ', 'Firstsource ', 'Flipkart ', 'Forbes & Company Limited ', 'Force Motors ', 'Fortis Healthcare ', 'Future Group ', 'GAI
L ', 'Galaxy Surfactants ', 'Garden Reach Shipbuilders & Engineers ', 'Geojit Financial Services ', 'Gini & Jony ', 'Glenmark Pharmaceuticals ', 'Glider
s India ', 'GMR Group ', 'Go First ', 'Goa Shipyard ', 'Godrej Group ', 'Great Eastern Shipping ', 'Greaves Cotton ', 'Greenply ', 'Gujarat Alkalies and
Chemicals Limited ', 'Gujarat Mineral Development Corporation ', 'Gujarat State Fertilizers and Chemicals ', 'Gujarat State Petroleum Corporation ', 'GV
K Industries ', "Haldiram's ", 'Haptik ', 'Hatsun Agro Product ', 'Havells ', 'Hawkins Cookers ', 'HCLTech ', 'HDFC Bank ', 'HDIL ', 'Hero Cycles ', 'He
ro FinCorp ', 'Hero MotoCorp ', 'HFCL ', 'Himalaya Wellness Company ', 'Hindalco Industries ', 'Hinduja Group ', 'Hinduja Healthcare ', 'Hindustan Aeron
autics Limited ', 'Hindustan Construction Company ', 'Hindustan Copper ', 'Hindustan Motors ', 'Hindustan Petroleum ', 'Hindustan Shipyard ', 'Hindustan
Times ', 'Hindustan Unilever ', 'Hindustan Zinc ', 'HLL Lifecare ', 'HMT Limited ', 'Housing Development Finance Corporation ', 'iball ', 'ICICI Bank ',
'IDBI Bank ', 'IDFC First Bank ', 'Indegene ', 'India Cements ', 'India Optel ', 'Indiabulls ', 'India Infoline ', 'Indian Bank ', 'Indian Hotels Compan
```

Accessing Industry Type of Companies

```
[14]: #Now, forming separate list of Industry type
      Industry_type=[]
      for i in range(1,len(new_list),8):
          Industry_type.append(new_list[i])
      print(Industry_type)
```

```
['Technology ', 'Energy ', 'Industrials ', 'Consumer services ', 'Industrials ', 'Industrials ', 'Insurance ', 'Industrials ', 'Conglomerate ', 'Conglom
erate ', 'Utilities ', 'Retail ', 'Conglomerate ', 'Conglomerate ', 'Industrials ', 'Industrials ', 'Insurance ', 'Consumer services ', 'Consumer servic
es ', 'Consumer services ', 'Conglomerate ', 'Health care ', 'Industrials ', 'Retail ', 'Industrials ', 'Health care ', 'Consumer goods ', 'Financials
', 'Health care ', 'Consumer goods ', 'Industrials ', 'Retail ', 'Consumer goods ', 'Industrials ', 'Health care ', 'Conglomerate ', 'Industrials ', 'In
dustrials ', 'Industrials ', 'Chemicals ', 'Consumer goods ', 'Conglomerate ', 'Financials ', 'Consumer goods ', 'Conglomerate ', 'Consumer services ',
'Consumer goods ', 'Basic materials ', 'Financials ', 'Financials ', 'Financials ', 'Industrials ', 'Industrials ', 'Industrials ', 'Basic materials ',
'Health care ', 'Industrials ', 'Industrials ', 'Industrials ', 'Industrials ', 'Energy ', 'Telecommunications ', 'Energy ', 'Telecommunications ', 'Hea
lth care ', 'Consumer goods ', 'Consumer goods ', 'Basic materials ', 'Consumer goods ', 'Industrials ', 'Consumer goods ', 'Industrials ', 'Real estate
', 'Consumer goods ', 'Consumer services ', 'Financials ', 'Retail ', 'Consumer goods ', 'Consumer goods ', 'Financials ', 'Industrials ', 'Industrials
', 'Utilities ', 'Industrials ', 'Energy ', 'Health care ', 'Financials ', 'Industrials ', 'Energy ', 'Chemicals ', 'Industrials ', 'Financials ', 'Indu
strials ', 'Chemicals ', 'Technology ', 'Retail ', 'Financials ', 'Technology ', 'Health care ', 'Utilities ', 'Technology ', 'Consumer services ', 'Ind
ustrials ', 'Basic materials ', 'Financials ', 'Consumer services ', 'Industrials ', 'Real estate ', 'Retail ', 'Health care ', 'Industrials ', 'Consume
r services ', 'Financials ', 'Industrials ', 'Consumer goods ', 'Industrials ', 'Business services ', 'Conglomerate ', 'Health care ', 'Industrials ',
'Basic materials ', 'Industrials ', 'Conglomerate ', 'Conglomerate ', 'Consumer goods ', 'Consumer goods ', 'Industrials ', 'Financials ', 'Financials
', 'Industrials ', 'Technology ', 'Conglomerate ', 'Consumer goods ', 'Health care ', 'Retail ', 'Utilities ', 'Chemicals ', 'Industrials ', 'Financials
', 'Consumer goods ', 'Health care ', 'Industrials ', 'Real estate ', 'Consumer services ', 'Industrials ', 'Conglomerate ', 'Industrials ', 'Industrial
s ', 'Industrials ', 'Chemicals ', 'Basic materials ', 'Chemicals ', 'Energy ', 'Conglomerate ', 'Consumer goods ', 'Technology ', 'Consumer goods ', 'I
ndustrials ', 'Consumer goods ', 'Industrials ', 'Financials ', 'Real estate ', 'Consumer goods ', 'Financials ', 'Consumer goods ', 'Telecommunications
', 'Health care ', 'Basic materials ', 'Conglomerate ', 'Health care ', 'Industrials ', 'Real estate ', 'Basic materials ', 'Consumer goods ', 'Energy
', 'Industrials ', 'Consumer services ', 'Consumer goods ', 'Basic materials ', 'Health care ', 'Industrials ', 'Financials ', 'Consumer goods ', 'Finan
cials ', 'Financials ', 'Financials ', 'Health care ', 'Industrials ', 'Industrials ', 'Financials ', 'Financials ', 'Financials ', 'Consumer services
', 'Energy ', 'Financials ', 'Financials ', 'Industrials ', 'Industrials ', 'Telecommunications ', 'Consumer services ', 'Financials ', 'Technology ',
'Industrials ', 'Financials ', 'Financials ', 'Technology ', 'Industrials ', 'Utilities ', 'Industrials ', 'Industrials ', 'Consumer goods ', 'Industria
ls ', 'Conglomerate ', 'Technology ', 'Technology ', 'Industrials ', 'Industrials ', 'Consumer goods ', 'Industrials ', 'Conglomerate ', 'Consumer servi
ces ', 'Energy ', 'Telecommunications ', 'Conglomerate ', 'Basic materials ', 'Basic materials ', 'Railways ', 'Utilities ', 'Conglomerate ', 'Consumer
```

Now, Accessing Industry Sector

```
[16]: Sector_type=[]
      for i in range(2,len(new_list),8):
          Sector_type.append(new_list[i])
      print(Sector_type)
```

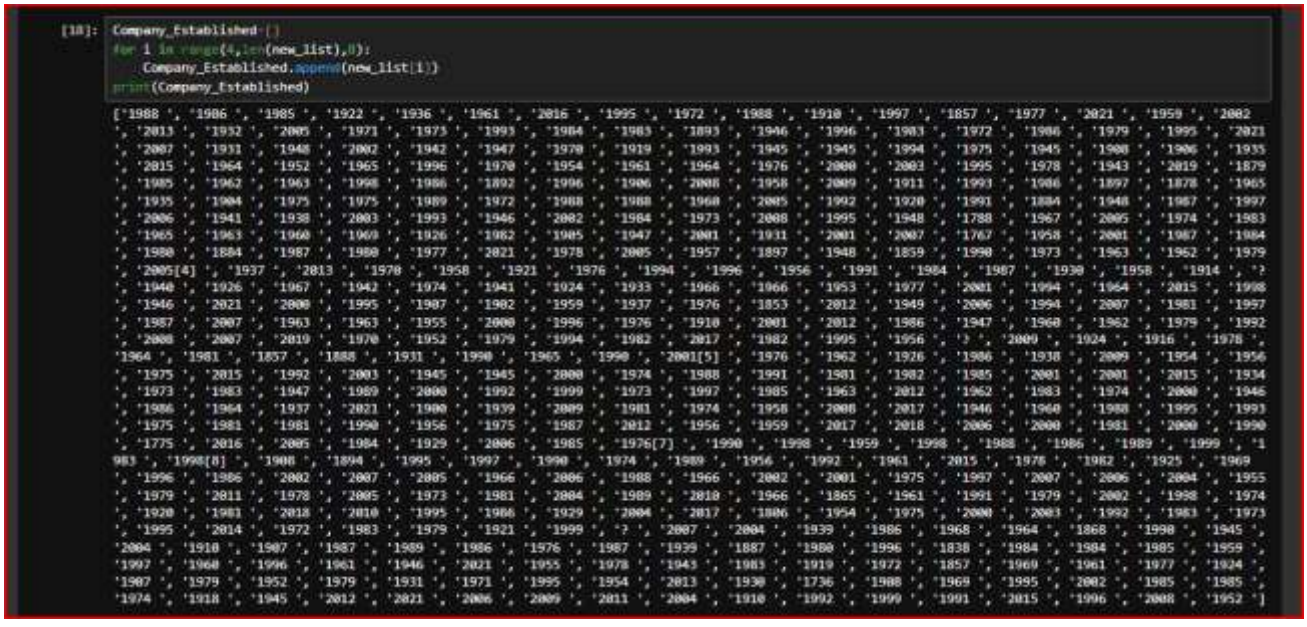
```
['Software ', 'Oil & gas offshore drilling & other services ', 'Shipbuilding ', 'Broadcasting & entertainment ', 'Cement ', 'Machinery: specialty ', 'Fu
ll line insurance ', 'Machinery: construction & handling ', '- ', '- ', 'Alternative electricity ', 'Apparel retailers ', '- ', '- ', 'Defense ', 'Const
ruction ', 'Full line insurance ', 'Airlines ', 'Airlines ', 'Airlines ', '- ', 'Pharmaceuticals ', 'Transportation services ', 'Apparel retailers ', 'C
ement ', 'Ayurvedic producers ', 'Food products ', 'Investment services ', 'Health care providers ', 'Tires ', 'Business training & employment agencies
', 'Specialty retailers ', 'Consumer electronics ', 'Defense ', 'Health care providers ', '- ', 'Commercial vehicles & parts ', 'Commercial vehicles & p
arts ', 'Paints and coatings ', 'Chemicals: diversified ', 'Automobiles ', '- ', 'Banks ', 'Automobiles ', '- ', 'Broadcasting & entertainment ', 'Food
products ', 'Paper ', 'Banks ', 'Banks ', 'Banks ', 'Aerospace, space industry ', 'Commercial vehicles & parts ', 'Engineering and contracting services
', 'Aluminium ', 'Pharmaceuticals & biotechnology ', 'Aerospace & defense ', 'Aerospace & defense ', 'Industrial engineering ', 'Electrical components
', 'Oil refining & marketing ', 'Telecommunications service providers ', 'Alternative energy ', 'Telecommunications services ', 'Pharmaceuticals & biote
chnology ', 'Consumer electronics ', 'Auto services ', 'Textile products ', 'Food products ', 'Glass ', 'Consumer electronics ', 'Aerospace & defense ',
'Real estate ', 'Food products ', 'Restaurants & bars ', 'Banks ', 'Specialty retailers ', 'Tires ', 'Consumer electronics ', 'Banks ', 'Electronic & el
ectrical equipment ', 'Building materials: other ', 'Alternative electricity ', 'Diversified industrials ', 'Oil refining & marketing ', 'Pharmaceutical
s ', 'Banks ', 'Business support services ', 'Coal ', 'Specialty chemicals ', 'Shipbuilding ', 'Investment services ', 'Railroads ', 'Fertilizers ', 'Se
miconductors ', 'Specialty retailers ', 'Banks ', 'Software & IT services ', 'Ayurvedic producers ', 'Conventional electricity ', 'Software ', 'Airlines
', 'Railroads ', 'General mining ', 'Banks ', 'Broadcasting & entertainment ', 'Electronic & electrical equipment ', 'Real estate holding & development
', 'Diversified retailers ', 'Pharmaceuticals ', 'Industrial engineering ', 'Travel & tourism ', 'Financial services ', 'Commercial vehicles & parts ',
'Food products ', 'Electronic & electrical equipment ', 'Entry-level employment ', '- ', 'Pharmaceuticals ', 'Engineering and contracting services ', 'G
eneral mining ', 'Industrial engineering ', '- ', '- ', 'Consumer electronics ', 'Durable household products ', 'Electrical components ', 'Life insuranc
e ', 'Banks ', 'Business support services ', 'Consumer digital services ', '- ', 'Automobiles ', 'Health care providers ', 'Diversified retailers ', 'Ga
s distribution ', 'Specialty chemicals ', 'Shipbuilding ', 'Investment services ', 'Clothing & accessories ', 'Pharmaceuticals ', 'Defense ', 'Real esta
te holding & development ', 'Airlines ', 'Shipbuilding ', '- ', 'Marine transportation ', 'Industrial engineering ', 'Building materials: other ', 'Spec
ialty chemicals ', 'General mining ', 'Fertilizers ', 'Oil refining & marketing ', '- ', 'Food products ', 'Software ', 'Food products ', 'Electrical co
mponents ', 'Durable household products ', 'Business support services ', 'Banks ', 'Real estate holding & development ', 'Recreational vehicles ', 'Fina
ncial services ', 'Automobiles ', 'Telecommunications equipment ', 'Ayurvedic producers ', 'Aluminium ', '- ', 'Health care providers ', 'Aerospace & de
fense ', 'Real estate holding & development ', 'Copper ', 'Automobiles ', 'Oil refining & marketing ', 'Shipbuilding ', 'Publishing ', 'Food & personal
products ', 'Nonferrous metals ', 'Pharmaceuticals & medical equipment ', 'Diversified industrials ', 'Financial services ', 'Consumer electronics ', 'B
```

Accessing Head Office Location

```
[17]: Head_Office=[]
for i in range(3,len(new_list),8):
    Head_Office.append(new_list[i])
print(Head_Office)
```

```
['Mumbai ', 'Chennai ', 'Ahmedabad ', 'Kolkata ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Faridabad ', 'New Delhi ', 'Ahmedabad ', 'Mumbai ', 'Mumbai ', 'Mumb
ai ', 'Mumbai ', 'Kanpur ', 'Mumbai ', 'New Delhi ', 'Vijayawada ', 'New Delhi ', 'Kochi ', 'Ahmedabad ', 'Mumbai ', 'Mumbai ', 'Gurugram ', 'Mumbai ',
'Chennai ', 'Anand ', 'Mumbai ', 'Chennai ', 'Gurugram ', 'Mumbai ', 'New Delhi ', 'Delhi ', 'Chennai ', 'Gurugram ', 'Ahmedabad ', 'Chennai ', 'Mumbai
', 'Mumbai ', 'Ahmedabad ', 'Rajkot ', 'New Delhi ', 'Ahmedabad ', 'Pune ', 'Pune ', 'Mumbai ', 'Mumbai ', 'Gurugram ', 'Vadodara ', 'Mumbai ', 'Pune ',
'Bengaluru ', 'Bengaluru ', 'Chennai ', 'New Delhi ', 'Hyderabad ', 'Hyderabad ', 'Bengaluru ', 'Pune ', 'New Delhi ', 'Mumbai ', 'New Delhi ', 'Chennai
', 'New Delhi ', 'Bengaluru ', 'Mumbai ', 'Gurugram ', 'Mumbai ', 'Ludhiana ', 'Mumbai ', 'Palakkad ', 'New Delhi ', 'Bangalore ', 'Kolkata ', 'Bengalur
u ', 'Bangalore ', 'Chennai ', 'Mumbai ', 'Hyderabad ', 'Mumbai ', 'Bangalore ', 'Kolkata ', 'Kolkata ', 'Mumbai ', 'Chennai ', 'Mumbai ', 'Kumbakonam
', 'New Delhi ', 'Kolkata ', 'Kochi ', 'Kochi ', 'Chennai ', 'New Delhi ', 'Hyderabad ', 'Bengaluru ', 'Mumbai ', 'Thrissur ', 'Hyderabad ', 'Ghaziabad
', 'Kolkata ', 'Mumbai ', 'Bengaluru ', 'New Delhi ', 'Panaji ', 'Mumbai ', 'Noida ', 'Noida ', 'New Delhi ', 'Mumbai ', 'Hyderabad ', 'Bangalore ', 'Ne
w Delhi ', 'Mumbai ', 'Gurugram ', 'Chennai ', 'Hyderabad ', 'Bangalore ', 'Kolkata ', 'Pune ', 'New Delhi ', 'New Delhi ', 'Faridabad ', 'Mumbai ', 'Mu
mbai ', 'Mumbai ', 'Kolkata ', 'Kolkata ', 'Bengaluru ', 'Kochi ', 'Mumbai ', 'Bengaluru ', 'Mumbai ', 'Pune ', 'Gurugram ', 'Mumbai ', 'New Delhi ', 'N
avi Mumbai ', 'Kolkata ', 'Kochi ', 'Mumbai ', 'Mumbai ', 'Nagpur ', 'New Delhi ', 'Mumbai ', 'Goa ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Tinsukia ', 'Noi
da ', 'Ahmedabad ', 'Vadodara ', 'Ahmedabad ', 'Hyderabad ', 'Bikaner ', 'Mumbai ', 'Chennai ', 'Noida ', 'Mumbai ', 'Noida ', 'Mumbai ', 'Mumbai ', 'Lu
dhiana ', 'New Delhi ', 'Gurugram ', 'New Delhi ', 'Bengaluru ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Bengaluru ', 'Mumbai ', 'Kolkata ', 'Kolkata ', 'Mumb
ai ', 'Visakhapatnam ', 'New Delhi ', 'Mumbai ', 'Udaipur ', 'Thiruvananthapuram ', 'Bengaluru ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ',
'Bangalore ', 'Tirunelveli ', 'Dehradun ', 'Gurugram ', 'Mumbai ', 'Chennai ', 'Mumbai ', 'New Delhi ', 'Chennai ', 'New Delhi ', 'New Delhi ', 'New Del
hi ', 'Bengaluru ', 'Gurugram ', 'Mumbai ', 'Ahmedabad ', 'Bengaluru ', 'Chennai ', 'Mumbai ', 'Bangalore ', 'Mumbai ', 'Noida ', 'Chennai ', 'Mumbai ',
'New Delhi ', 'New Delhi ', 'Kolkata ', 'Bengaluru ', 'Gurugram ', 'Jalgaon ', 'Noida ', 'Mysuru ', 'Bengaluru ', 'Noida ', 'Mumbai ', 'Ahmedabad ', 'Na
vi Mumbai ', 'Navi Mumbai ', 'New Delhi ', 'New Delhi ', 'Kolkata ', 'Mumbai ', 'Mumbai ', 'Gurugram ', 'Mumbai ', 'Noida ', 'Mumbai ', 'Pune ', 'New De
lhi ', 'Mangaluru ', 'Karur ', 'Thiruvananthapuram ', 'Kochi ', 'Kolkata ', 'Bangaluru ', 'Pune ', 'Mumbai ', 'Navi Mumbai ', 'Mumbai ', 'Pune ', 'Hyder
abad ', 'Mangaluru ', 'Coimbatore ', 'Chennai ', 'New Delhi ', 'Mumbai ', 'Noida ', 'Karnal ', 'Mumbai ', 'New Delhi ', 'Mumbai ', 'Sambalpur ', 'Mumbai
', 'Mumbai ', 'Mumbai ', 'Gurugram ', 'Bengaluru ', 'Mangaluru ', 'Mumbai ', 'New Delhi ', 'Mumbai ', 'New Delhi ', 'New Delhi ', 'New Delhi ', 'New Del
hi ', 'Mumbai ', 'Ranchi ', 'Mumbai ', 'Mumbai ', 'Bengaluru ', 'Gurugram ', 'Erode ', 'Bengaluru ', 'Hyderabad ', 'Bengaluru ', 'Kanpur ', 'New Delhi
', 'Raebareli ', 'Nagpur ', 'New Delhi ', 'Noida ', 'Bengaluru ', 'Chennai ', 'Delhi ', 'Kolkata ', 'Mumbai ', 'Kanpur ', 'Chennai ', 'Kochi ', 'Bengalu
ru ', 'Bhubaneswar ', 'Noida ', 'Hyderabad ', 'Mumbai ', 'Mumbai ', 'Jaipur ', 'New Delhi ', 'New Delhi ', 'Chandigarh ', 'Noida ', 'Faridabad ', 'Gurug
ram ', 'Mumbai ', 'Ahmedabad ', 'Neyveli ', 'New Delhi ', 'Mumbai ', 'Mumbai ', 'Dehradun ', 'Duliajan ', 'Bangaluru ', 'New Delhi ', 'Mangalore ', 'Noi
da ', 'Mumbai ', 'Bengaluru ', 'Mumbai ', 'Kolkata ', 'New Delhi ', 'Chennai ', 'Mumbai ', 'Mumbai ', 'Haridwar ', 'Noida ', 'Chennai[7] ', 'Nagpur ',
'New Delhi ', 'Mumbai ', 'Noida ', 'Mumbai ', 'New Delhi ', 'Gurugram ', 'New Delhi ', 'Pune ', 'New Delhi ', 'New Delhi ', 'New Delhi ', 'Mumbai ', 'Gu
rugram ', 'New Delhi ', 'Gurugram ', 'Bengaluru ', 'Mumbai ', 'Chennai ', 'Gurugram ', 'Bengaluru ', 'Mumbai ', 'Visakhapatnam ', 'Mumbai ', 'New Delhi
', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Mumbai ', 'Rajula ', 'Navi
```

Accessing Establishment Year of company



```
[18]: Company_Established=[]
for i in range(4,len(new_list),8):
    Company_Established.append(new_list[i])
print(Company_Established)
```

```
['1988 ', '1986 ', '1985 ', '1922 ', '1936 ', '1961 ', '2016 ', '1995 ', '1972 ', '1988 ', '1910 ', '1997 ', '1857 ', '1977 ', '2021 ', '1959 ', '2002
', '2013 ', '1932 ', '2005 ', '1971 ', '1973 ', '1993 ', '1984 ', '1983 ', '1893 ', '1946 ', '1996 ', '1983 ', '1972 ', '1986 ', '1979 ', '1995 ', '2021
', '2007 ', '1931 ', '1948 ', '2002 ', '1942 ', '1947 ', '1970 ', '1919 ', '1993 ', '1945 ', '1945 ', '1994 ', '1975 ', '1945 ', '1980 ', '1906 ', '1935
', '2015 ', '1964 ', '1952 ', '1965 ', '1996 ', '1970 ', '1954 ', '1961 ', '1964 ', '1976 ', '2000 ', '2003 ', '1995 ', '1978 ', '1943 ', '2019 ', '1879
', '1985 ', '1962 ', '1963 ', '1998 ', '1986 ', '1892 ', '1996 ', '1906 ', '2008 ', '1958 ', '2009 ', '1911 ', '1993 ', '1986 ', '1897 ', '1878 ', '1965
', '1935 ', '1984 ', '1975 ', '1975 ', '1989 ', '1972 ', '1988 ', '1988 ', '1960 ', '2005 ', '1992 ', '1920 ', '1991 ', '1884 ', '1948 ', '1987 ', '1997
', '2006 ', '1941 ', '1938 ', '2003 ', '1993 ', '1946 ', '2002 ', '1984 ', '1973 ', '2008 ', '1995 ', '1948 ', '1788 ', '1967 ', '2005 ', '1974 ', '1983
', '1965 ', '1963 ', '1960 ', '1969 ', '1926 ', '1982 ', '1985 ', '1947 ', '2001 ', '1931 ', '2001 ', '2007 ', '1767 ', '1958 ', '2001 ', '1987 ', '1984
', '1980 ', '1884 ', '1987 ', '1980 ', '1977 ', '2021 ', '1978 ', '2005 ', '1957 ', '1897 ', '1948 ', '1859 ', '1990 ', '1973 ', '1963 ', '1962 ', '1979
', '2005[4] ', '1937 ', '2013 ', '1978 ', '1958 ', '1921 ', '1976 ', '1994 ', '1996 ', '1956 ', '1991 ', '1984 ', '1987 ', '1930 ', '1958 ', '1914 ', '?
', '1940 ', '1926 ', '1967 ', '1942 ', '1974 ', '1941 ', '1924 ', '1933 ', '1966 ', '1966 ', '1953 ', '1977 ', '2001 ', '1994 ', '1964 ', '2015 ', '1988
', '1946 ', '2021 ', '2000 ', '1995 ', '1987 ', '1982 ', '1959 ', '1937 ', '1976 ', '1853 ', '2012 ', '1949 ', '2006 ', '1994 ', '2007 ', '1981 ', '1957
', '1987 ', '2007 ', '1963 ', '1963 ', '1955 ', '2000 ', '1996 ', '1976 ', '1910 ', '2001 ', '2012 ', '1986 ', '1947 ', '1960 ', '1962 ', '1979 ', '1992
', '2008 ', '2007 ', '2019 ', '1970 ', '1952 ', '1979 ', '1994 ', '1982 ', '2017 ', '1982 ', '1995 ', '1956 ', '? ', '2009 ', '1924 ', '1916 ', '1978 ',
'1964 ', '1981 ', '1857 ', '1888 ', '1931 ', '1990 ', '1965 ', '1990 ', '2001[5] ', '1976 ', '1962 ', '1926 ', '1986 ', '1938 ', '2009 ', '1954 ', '1956
', '1975 ', '2015 ', '1992 ', '2003 ', '1945 ', '1945 ', '2000 ', '1974 ', '1988 ', '1991 ', '1981 ', '1982 ', '1985 ', '2001 ', '2001 ', '2015 ', '1934
', '1973 ', '1983 ', '1947 ', '1989 ', '2000 ', '1992 ', '1999 ', '1973 ', '1997 ', '1985 ', '1963 ', '2012 ', '1962 ', '1983 ', '1974 ', '2000 ', '1946
', '1986 ', '1964 ', '1937 ', '2021 ', '1900 ', '1939 ', '2009 ', '1981 ', '1974 ', '1958 ', '2008 ', '2017 ', '1946 ', '1960 ', '1988 ', '1995 ', '1993
', '1975 ', '1981 ', '1981 ', '1990 ', '1956 ', '1975 ', '1987 ', '2012 ', '1956 ', '1959 ', '2017 ', '2018 ', '2006 ', '2000 ', '1981 ', '2000 ', '1990
', '1775 ', '2016 ', '2005 ', '1984 ', '1929 ', '2006 ', '1985 ', '1976[7] ', '1990 ', '1998 ', '1959 ', '1998 ', '1988 ', '1986 ', '1989 ', '1999 ', '1
983 ', '1998[8] ', '1900 ', '1894 ', '1995 ', '1997 ', '1990 ', '1974 ', '1989 ', '1956 ', '1992 ', '1961 ', '2015 ', '1978 ', '1982 ', '1925 ', '1969
', '1996 ', '1986 ', '2002 ', '2007 ', '2005 ', '1966 ', '2006 ', '1988 ', '1966 ', '2002 ', '2001 ', '1975 ', '1997 ', '2007 ', '2006 ', '2004 ', '1955
', '1979 ', '2011 ', '1978 ', '2005 ', '1973 ', '1981 ', '2004 ', '1989 ', '2010 ', '1966 ', '1865 ', '1961 ', '1991 ', '1979 ', '2002 ', '1998 ', '1974
', '1920 ', '1981 ', '2018 ', '2010 ', '1995 ', '1986 ', '1929 ', '2004 ', '2017 ', '1886 ', '1954 ', '1975 ', '2000 ', '2003 ', '1992 ', '1983 ', '1973
', '1995 ', '2014 ', '1972 ', '1983 ', '1979 ', '1921 ', '1999 ', '? ', '2007 ', '2004 ', '1939 ', '1986 ', '1968 ', '1964 ', '1868 ', '1998 ', '1945 ',
'2004 ', '1910 ', '1907 ', '1987 ', '1989 ', '1986 ', '1976 ', '1987 ', '1939 ', '1887 ', '1980 ', '1996 ', '1838 ', '1984 ', '1984 ', '1985 ', '1959 ',
'1997 ', '1968 ', '1996 ', '1961 ', '1946 ', '2021 ', '1955 ', '1978 ', '1943 ', '1983 ', '1919 ', '1972 ', '1857 ', '1969 ', '1961 ', '1977 ', '1924 ',
'1907 ', '1979 ', '1952 ', '1979 ', '1931 ', '1971 ', '1995 ', '1954 ', '2013 ', '1930 ', '1736 ', '1908 ', '1969 ', '1995 ', '2002 ', '1985 ', '1985 ',
'1974 ', '1918 ', '1945 ', '2012 ', '2021 ', '2006 ', '2009 ', '2011 ', '2004 ', '1910 ', '1992 ', '1999 ', '1991 ', '2015 ', '1996 ', '2008 ', '1952 ']
```

Using Website Scrapping and Pandas library

```
[22]: import pandas as pd

df=pd.DataFrame({"Company":Company_Names,"Industry":Industry_type,"Sector":Sector_type,"Main_Office":Head_Office,"start_up":Company_Established,
                 })
df
```

[22]:

	Company	Industry	Sector	Main_Office	start_up
0	63 Moons Technologies	Technology	Software	Mumbai	1988
1	Aban Offshore	Energy	Oil & gas offshore drilling & other services	Chennai	1986
2	ABG Shipyard	Industrials	Shipbuilding	Ahmedabad	1985
3	ABP Group	Consumer services	Broadcasting & entertainment	Kolkata	1922
4	ACC	Industrials	Cement	Mumbai	1936
...	...	...	...	...	...
487	Zensar Technologies	Technology	Software & IT services	Pune	1991
488	Zeta	Technology	Consumer digital services	Mumbai	2015
489	Zoho Corporation	Technology	Software	Chennai	1996
490	Zomato	Technology	Consumer digital services	Gurugram	2008
491	Zydus Lifesciences	Health care	Pharmaceuticals	Ahmedabad	1952

492 rows × 5 columns

Exporting all the data into a CSV File

```
[24]: df.to_csv("Top Companies Web Scrapping python.csv")

[25]: pwd()

[25]: 'C:\\Users\\DELL\\Downloads'

[26]: read=pd.read_csv("Top Companies Web Scrapping python.csv")
      read
```

[26]:

	Unnamed: 0	Company	Industry	Sector	Main_Office	start_up
0	0	63 Moons Technologies	Technology	Software	Mumbai	1988
1	1	Aban Offshore	Energy	Oil & gas offshore drilling & other services	Chennai	1986
2	2	ABG Shipyard	Industrials	Shipbuilding	Ahmedabad	1985
3	3	ABP Group	Consumer services	Broadcasting & entertainment	Kolkata	1922
4	4	ACC	Industrials	Cement	Mumbai	1936
...	...	...	...	...	...	...
487	487	Zensar Technologies	Technology	Software & IT services	Pune	1991
488	488	Zeta	Technology	Consumer digital services	Mumbai	2015
489	489	Zoho Corporation	Technology	Software	Chennai	1996
490	490	Zomato	Technology	Consumer digital services	Gurugram	2008
491	491	Zydus Lifesciences	Health care	Pharmaceuticals	Ahmedabad	1952

492 rows × 6 columns

Conclusion:

The web scraping project focused on identifying and analyzing the top companies in India has been a valuable learning experience in leveraging Python for data collection and processing. By utilizing the BeautifulSoup library, we successfully extracted detailed information about companies, including their rankings and industry sectors. This project highlights the power of web scraping for gathering real-time, publicly available data from online sources, making it a crucial skill in data science and analytics.

The data extracted was processed, cleaned, and analyzed to identify patterns and trends among top-performing companies. The generation of a professional PDF report added structure and readability to the findings, showcasing the data through well-organized tables and visualizations. This not only demonstrates technical skills in Python but also effective communication of data-driven insights, which is essential for decision-making.

Overall, this project emphasizes the importance of combining technical expertise with analytical thinking to transform raw data into meaningful insights. It also underscores the growing relevance of web scraping in understanding business landscapes and industry trends, making it a practical and impactful addition to a data-driven portfolio.

Om Suryawanshi