Web Scraping of AmbitionBox website

In [1]:
```
import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup
```

In [2]:
```
requests.get('https://www.ambitionbox.com/list-of-companies?page=1')
```

Out[2]:
```
<Response [403]>
```

In [3]:
```
headers={'User-Agent':'Mozilla/5.0 (Windows NT 6.3; Win 64 ; x64) Apple
```

In [4]:
```
webpage= requests.get('https://www.ambitionbox.com/list-of-companies?pa
```

In [5]:
```
webpage
```

Out[5]:
```
'<!doctype html>\n<html data-n-head-ssr lang="en" data-n-head="%7
B%22lang%22:%7B%22ssr%22:%22en%22%7D%7D">\n  <head >\n    <meta ch
arset="UTF-8">\n    <meta name="viewport" content="width=device-wi
dth,initial-scale=1,minimum-scale=1">\n    <meta http-equiv="X-UA-
Compatible" content="IE=edge"> \n    <title>List of companies - 79
0k companies | AmbitionBox</title><meta data-n-head="ssr" name="co
pyright" content="2023 AmbitionBox"><meta data-n-head="ssr" name
="revisit-after" content="1 day"><meta data-n-head="ssr" name="app
lication-name" content="AmbitionBox"><meta data-n-head="ssr" name
="content-language" content="EN"><meta data-n-head="ssr" name="goo
gle-signin-client_id" content="462822053404-hphug4pkahqljh2tc96g35
at47o4isv2.apps.googleusercontent.com"><meta data-n-head="ssr" pro
perty="fb:app_id" content="712617688793459"><meta data-n-head="ss
r" name="theme-color" content="#ffffff"><meta data-n-head="ssr" na
me="msapplication-navbutton-color" content="#ffffff"><meta data-n-
head="ssr" name="apple-mobile-web-app-status-bar-style" content="d
efault"><meta data-n-head="ssr" property="fb:admins" content="1000
01438127755,100000444923785"><meta data-n-head="ssr" property="og:
site_name" content="AmbitionBox"><meta data-n-head="ssr" property
```

In [6]:
```
soup = BeautifulSoup(webpage, 'lxml')
soup
```

Out[6]:
```
<!DOCTYPE html>
<html data-n-head="%7B%22lang%22:%7B%22ssr%22:%22en%22%7D%7D" data
-n-head-ssr="" lang="en">
<head>
<meta charset="utf-8"/>
<meta content="width=device-width,initial-scale=1,minimum-scale=1"
name="viewport"/>
<meta content="IE=edge" http-equiv="X-UA-Compatible"/>
<title>List of companies - 790k companies | AmbitionBox</title><me
ta content="2023 AmbitionBox" data-n-head="ssr" name="copyright"/>
<meta content="1 day" data-n-head="ssr" name="revisit-after"/><met
a content="AmbitionBox" data-n-head="ssr" name="application-name"/
><meta content="EN" data-n-head="ssr" name="content-language"/><me
ta content="462822053404-hphug4pkahqljh2tc96g35at47o4isv2.apps.goo
gleusercontent.com" data-n-head="ssr" name="google-signin-client_i
d"/><meta content="712617688793459" data-n-head="ssr" property="f
b:app_id"/><meta content="#ffffff" data-n-head="ssr" name="theme-c
olor"/><meta content="#ffffff" data-n-head="ssr" name="msapplicati
on-navbutton-color"/><meta content="default" data-n-head="ssr" nam
```

In [7]:
```
soup.find_all('h1')[0].text
```

Out[7]:
```
'List of companies in India'
```

In [8]:
```
for i in soup.find_all('h2'):
    print(i.text.strip())
```

```
TCS
Accenture
Cognizant
ICICI Bank
HDFC Bank
Wipro
Infosys
Capgemini
Tech Mahindra
Genpact
HCLTech
Amazon
Axis Bank
Concentrix Corpo...
IBM
Reliance jio
Larsen & Toubro ...
HDB Financial Se...
Vodafone Idea
Teleperformance
Reliance Retail
Kotak Mahindra B...
Reliance Industr...
Deloitte
Bharti Airtel
BYJU'S
Tata Motors
Flipkart
WNS
IndusInd Bank
```

In [9]:
```
company = soup.find_all('div', class_ = 'company-content-wrapper')
```

In [10]:
```
company
```

Out[10]:
```
[<div class="company-content-wrapper"><div class="company-conten
t"><div class="company-logo"><img alt="Tata Consultancy Services l
ogo" class="lazy" data-src="https://static.ambitionbox.com/assets/
v2/images/rs:fit:200:200:false:false/bG9jYWw6Ly8vbG9nb3Mvb3JpZ2luY
WxzL3Rjcy5qcGc.webp" height="100" onerror="this.onerror=null;this.
src='/static/icons/company-placeholder.svg';" src="https://static.
ambitionbox.com/static/icons/company-placeholder.svg" width="100"/
></div> <div class="company-info-wrapper"><div class="company-inf
o"><div class="left"><a href="/overview/tcs-overview"><h2 class="c
ompany-name bold-title-l" title="TCS">

TCS
                                                                </
h2></a> <div class="rating-wrapper"><p class="rating badge-large r
ating-35"><i class="icon icon-star"></i>

3.9
                                                                </
p> <a class="review-count sbold-Labels" href="https://www.ambition
```

In [11]:
```
len(company)
```

Out[11]: 30

In [12]:
```
name = []
rating = []
reviews = []
type_of_comp = []
headquarter = []
how_old = []
no_of_emp = []


for i in company:
    name.append(i.find('h2').text.strip())
    rating.append(i.find('p', class_='rating').text.strip())
    reviews.append(i.find('a', class_='review-count').text.strip())
    type_of_comp.append(i.find_all('p', class_='infoEntity')[0].text.st
    headquarter.append(i.find_all('p', class_='infoEntity')[1].text.str
    how_old.append(i.find_all('p', class_='infoEntity')[2].text.strip()
    no_of_emp.append(i.find_all('p', class_='infoEntity')[3].text.strip

df = pd.DataFrame({'Company Name': name, 'Company Rating' : rating, 'Co
                   'type_of_comp': type_of_comp, 'headquarter' : headqua
                   'no_of_emp': no_of_emp})
```

In [13]:
```
df
```

Out[13]:

	Company Name	Company Rating	Company Reviews	type_of_comp	headquarter
0	TCS	3.9	(53.3k Reviews)	Public	Mumbai,Maharashtra + 275 more
1	Accenture	4.1	(33.8k Reviews)	Public	Dublin + 133 more
2	Cognizant	4.0	(31.1k Reviews)	Private	Teaneck. New Jersey. + 102 more
3	ICICI Bank	4.0	(35k Reviews)	Public	Mumbai,Maharashtra + 1072 more
4	HDFC Bank	4.0	(41.3k Reviews)	Public	Mumbai,Maharashtra + 1295 more
5	Wipro	3.9	(30.4k Reviews)	Public	Bangalore/Bengaluru,Karnataka + 229 more
6	Infosys	3.9	(29.6k Reviews)	Public	Bengaluru/Bangalore,Karnataka + 127 more
7	Capgemini	3.9	(25.1k Reviews)	Public	Paris + 79 more
8	Tech Mahindra	3.7	(22.6k Reviews)	Public	Pune,Maharashtra + 213 more
9	Genpact	4.0	(20.3k Reviews)	Public	New York,New York + 69 more
10	HCLTech	3.8	(19.5k Reviews)	Public	Noida,Uttar Pradesh + 141 more
11	Amazon	4.2	(19.8k Reviews)	Public	Seattle,Washington + 387 more
12	Axis Bank	3.9	(19.3k Reviews)	Public	Mumbai,Maharashtra + 1123 more
13	Concentrix Corpo...	4.0	(15.6k Reviews)	Public	Fremont,California + 70 more
14	IBM	4.2	(15.8k Reviews)	Public	Armonk,New York + 121 more
15	Reliance jio	4.0	(15.4k Reviews)	Public	Navi Mumbai,Maharashtra + 1019 more
16	Larsen & Toubro ...	4.1	(25.1k Reviews)	Public	Mumbai,Maharashtra + 536 more
17	HDB Financial Se...	4.0	(13.9k Reviews)	Private	Ahmedabad,Gujrat + 812 more

	Company Name	Company Rating	Company Reviews	type_of_comp	headquarter
18	Vodafone Idea	4.2	(13.5k Reviews)	Public	Gandhinagar,Gujrat + 578 more
19	Teleperformance	3.6	(15.1k Reviews)	Private	Paris + 122 more
20	Reliance Retail	4.1	(17.8k Reviews)	Private	Navi Mumbai,Maharashtra + 739 more
21	Kotak Mahindra B...	3.9	(14k Reviews)	Public	Mumbai,Maharashtra + 481 more
22	Reliance Industr...	4.1	(46.9k Reviews)	Public	Navi Mumbai,Maharashtra + 519 more
23	Deloitte	4.1	(10.8k Reviews)	Private	New York,New York + 132 more
24	Bharti Airtel	4.1	(12.8k Reviews)	Public	Gurgaon/Gurugram,Haryana + 540 more
25	BYJU'S	3.4	(14k Reviews)	Private	Bangalore,Karnataka + 264 more
26	Tata Motors	4.1	(12.3k Reviews)	Public	Pune,Maharashtra + 394 more
27	Flipkart	4.2	(12.5k Reviews)	Public	Bangalore,Karnataka + 462 more
28	WNS	3.7	(7.2k Reviews)	Private	Mumbai,Maharashtra + 24 more
29	IndusInd Bank	3.8	(6.9k Reviews)	Public	Gurgaon/Gurugram,Haryana + 578 more

In [14]:
```
df.shape
```

Out[14]: (30, 7)

Extract all details from all 333 pages

In [15]:

```
final = pd.DataFrame()

for j in range(1,334):
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 6.3; Win 64 ; x64) A
    url = requests.get('https://www.ambitionbox.com/list-of-companies?p
    webpage= requests.get('https://www.ambitionbox.com/list-of-companie
    soup = BeautifulSoup(webpage, 'lxml')
    company = soup.find_all('div', class_ = 'company-content-wrapper')

    name = []
    rating = []
    reviews = []
    type_of_comp = []
    headquarter = []
    how_old = []
    no_of_emp = []

    for i in company:
        try:
            name.append(i.find('h2').text.strip())
        except:
            name.append(np.nan)
        try:
            rating.append(i.find('p', class_='rating').text.strip())
        except:
            rating.append(np.nan)
        try:
            reviews.append(i.find('a', class_='review-count').text.stri
        except:
            reviews.append(np.nan)
        try:
            type_of_comp.append(i.find_all('p', class_='infoEntity')[0]
        except:
            type_of_comp.append(np.nan)
        try:
            headquarter.append(i.find_all('p', class_='infoEntity')[1].
        except:
            headquarter.append(np.nan)
        try:
            how_old.append(i.find_all('p', class_='infoEntity')[2].text
        except:
            how_old.append(np.nan)
        try:
            no_of_emp.append(i.find_all('p', class_='infoEntity')[3].te
        except:
            no_of_emp.append(np.nan)

    df = pd.DataFrame({'Company Name': name, 'Company Rating' : rating,
                       'type_of_comp': type_of_comp, 'headquarter' : headqua
                       'no_of_emp': no_of_emp})

    final = final.append(df, ignore_index=True)
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_19552\825628277.py:52: F
utureWarning: The frame.append method is deprecated and will be remov
ed from pandas in a future version. Use pandas.concat instead.
  final = final.append(df, ignore_index=True)
```

In [16]:
```
final.tail()
```

Out[16]:

	Company Name	Company Rating	Company Reviews	type_of_comp	headquarter	how_old
9979	St. Andrews Inst...	4.5	(53 Reviews)	Gurgaon/Gurugram + 2 more	11-50 Employees (India)	NaN
9980	Sarla Advantech	4.1	(53 Reviews)	Private	Navi Mumbai,Maharashtra + 5 more	24 years old
9981	Pathways Schools	4.4	(53 Reviews)	Gurgaon/Gurugram + 2 more	NaN	NaN
9982	Vega Industries	4.1	(53 Reviews)	Private	Noida,Uttar Pradesh + 9 more	23 years old
9983	Tahal Group	4.3	(53 Reviews)	Or Yehuda + 7 more	71 years old	501-1I Employees (Global

In [17]:
```
final.shape
```

Out[17]:
```
(9984, 7)
```

In [ ]:
```

```