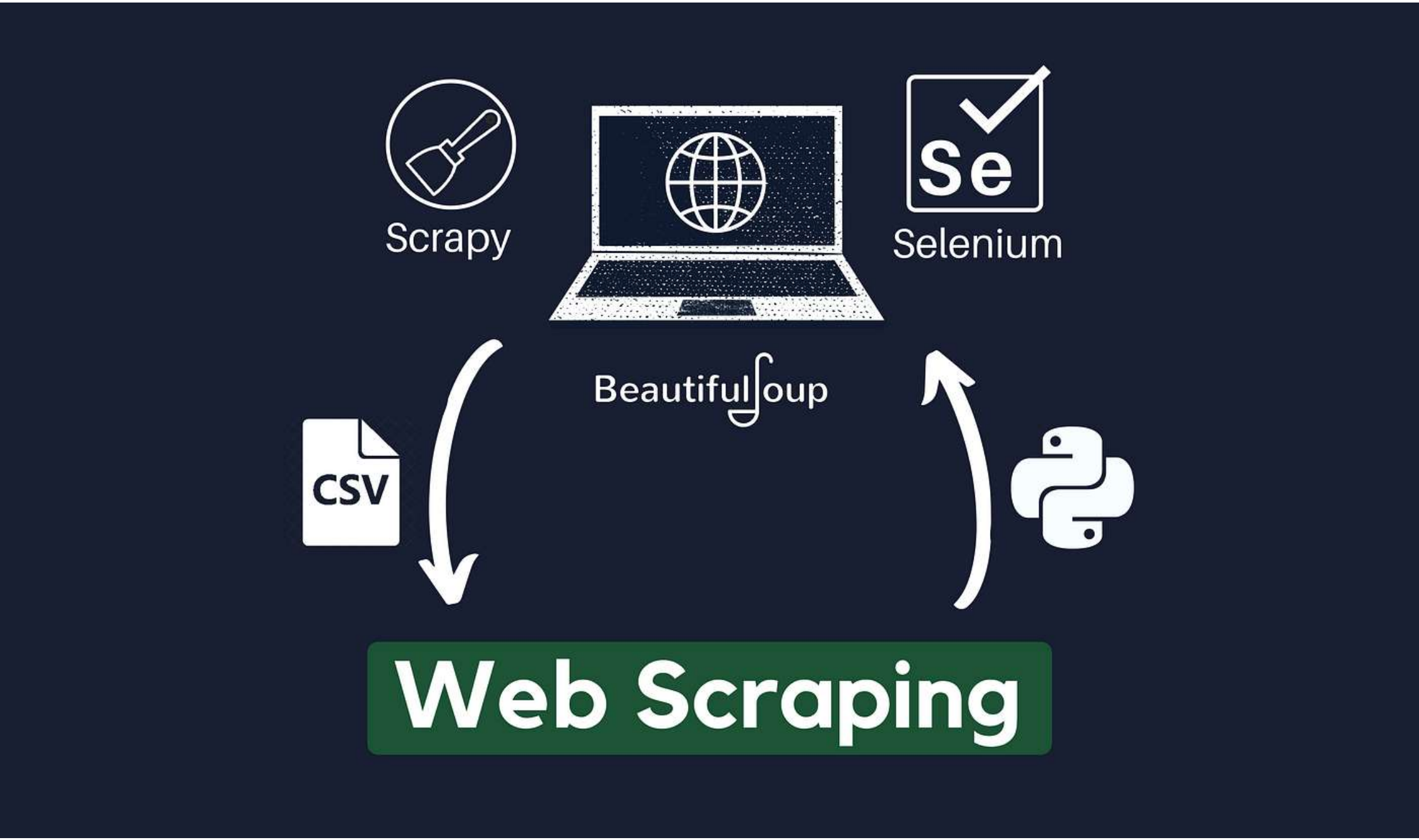

Scrapy
Se
Selenium
BeautifulSoup
CSV
Web Scraping

jupyter Web Scrapping from Wikipedia Last Checkpoint: 40 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [1]:
```
from bs4 import BeautifulSoup
import requests
```

In [2]:
```
url = "https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States_by_revenue"
page = requests.get(url)
soup = BeautifulSoup(page.text, 'html')
soup
```

Out[2]:
```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vecto
r-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-feat
ure-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled vector-f
eature-custom-font-size-clientpref-0 vector-feature-client-preferences-disabled vector-feature-client-prefs-pinned-disabled v
ector-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-h
eader-disabled vector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clie
ntpref-1 vector-feature-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-cont
ent-enabled vector-feature-custom-font-size-clientpref-0 vector-feature-client-preferences-disabled vector-feature-client-pre
fs-pinned-disabled vector-toc-available";var cookie=document.cookie.match(/(?:^|; )enwikimwclientpreferences=([^;]+)/);if(coo
kie){cookie[1].split('%2C').forEach(function(pref){className=className.replace(new RegExp('(^| )'+pref.replace(/-clientpref-
```

In [4]:
```
table = soup.find_all('table')[1]
table
```

Out[4]:
```
<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 [5]:
```
# world_titles = table.find_all('th')
world_titles = table.find_all('th')
```

In [7]:
```
world_table_titles = [title.text.strip() for title in world_titles]
world_table_titles
```

Out[7]:
```
['Rank',
 'Name',
 'Industry',
 'Revenue (USD millions)',
 'Revenue growth',
 'Employees',
 'Headquarters']
```

In [8]:
```
column_data = table.find_all('tr')
```

In [9]:
```
import pandas as pd
```

n [10]:
```
df = pd.DataFrame(columns = world_table_titles)
df
```

In [12]:
```
for row in column_data[1:]:
    row_data = row.find_all('td')
    individual_row_data = [data.text.strip() for data in row_data]
#     print(individual_row_data)
    length = len(df)
    df.loc[length] = individual_row_data
```

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

Out[13]:

	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 [118]:
```
from matplotlib import pyplot as plt
import numpy as np
```

In [119]:
```
%matplotlib inline
```

In [120]:
```
df=  df3.copy()
```

In [121]:
```
df=df.set_index('Rank')
```

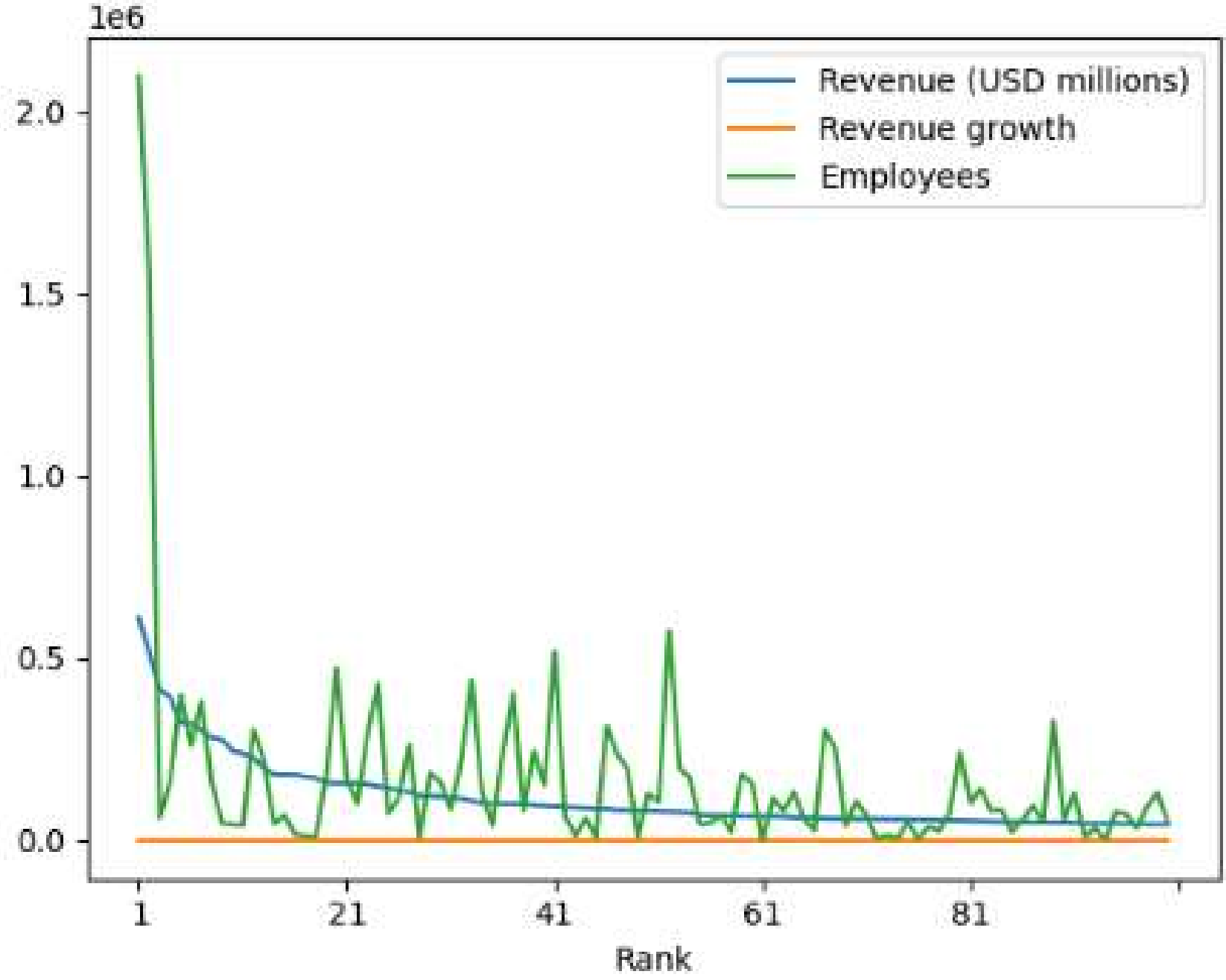
In [122]:
```
df.head()
```

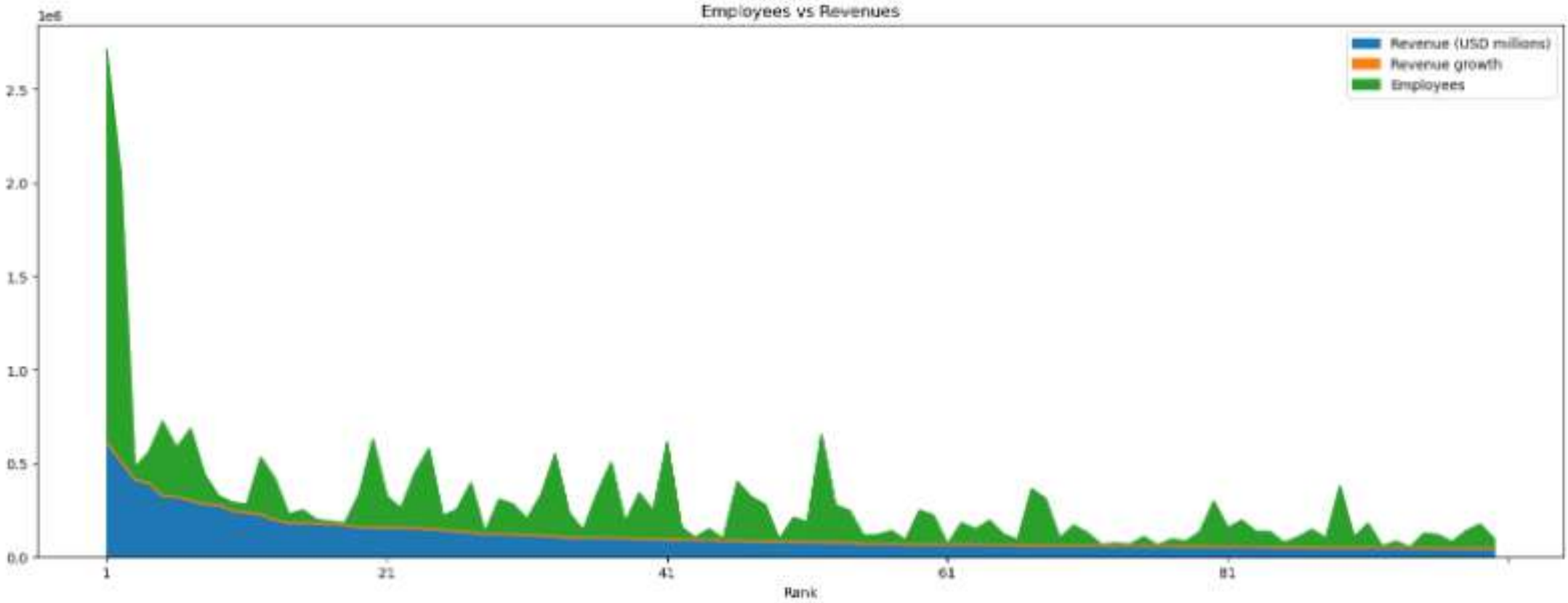
Out[122]:

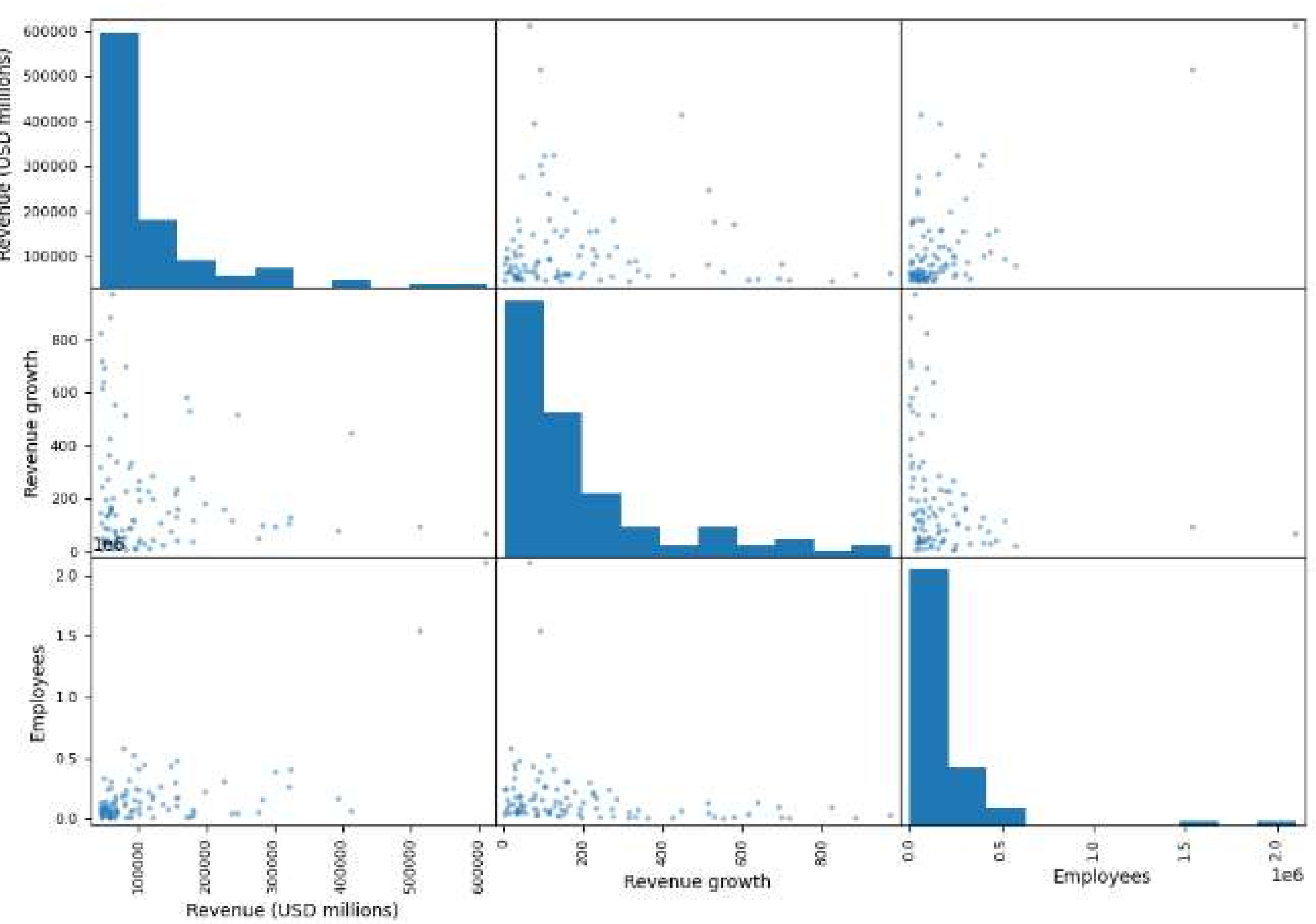
Rank	Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters
1	Walmart	Retail	611289	67	2100000	Bentonville, Arkansas
2	Amazon	Retail and cloud computing	513983	94	1540000	Seattle, Washington
3	ExxonMobil	Petroleum industry	413680	448	62000	Spring, Texas
4	Apple	Electronics industry	394328	78	164000	Cupertino, California
5	UnitedHealth Group	Healthcare	324162	127	400000	Minnetonka, Minnesota

```
In [123]: df.plot()
```

```
Out[123]: <Axes: xlabel='Rank'>
```




Employees vs Revenues
1e6
2.5
2.0
1.5
1.0
0.5
0.0
1
21
41
61
81
Rank
Revenue (USD millions)
Revenue growth
Employees


Revenue (USD millions)
Revenue growth
Employees
1e6