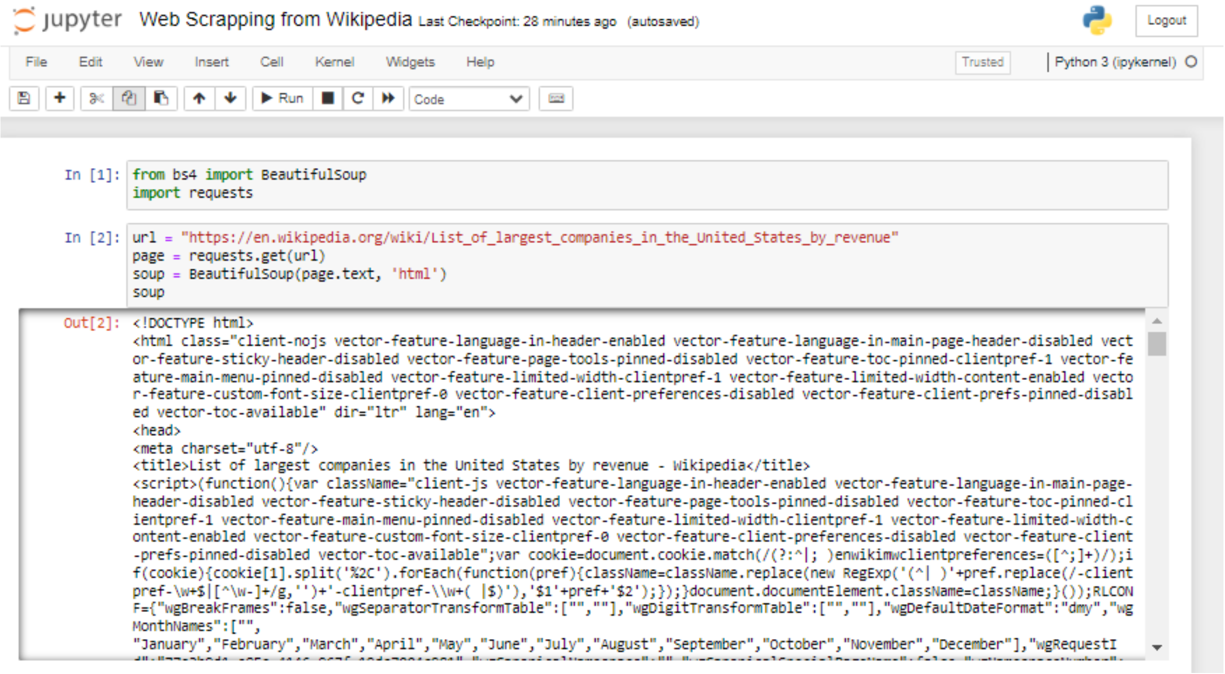
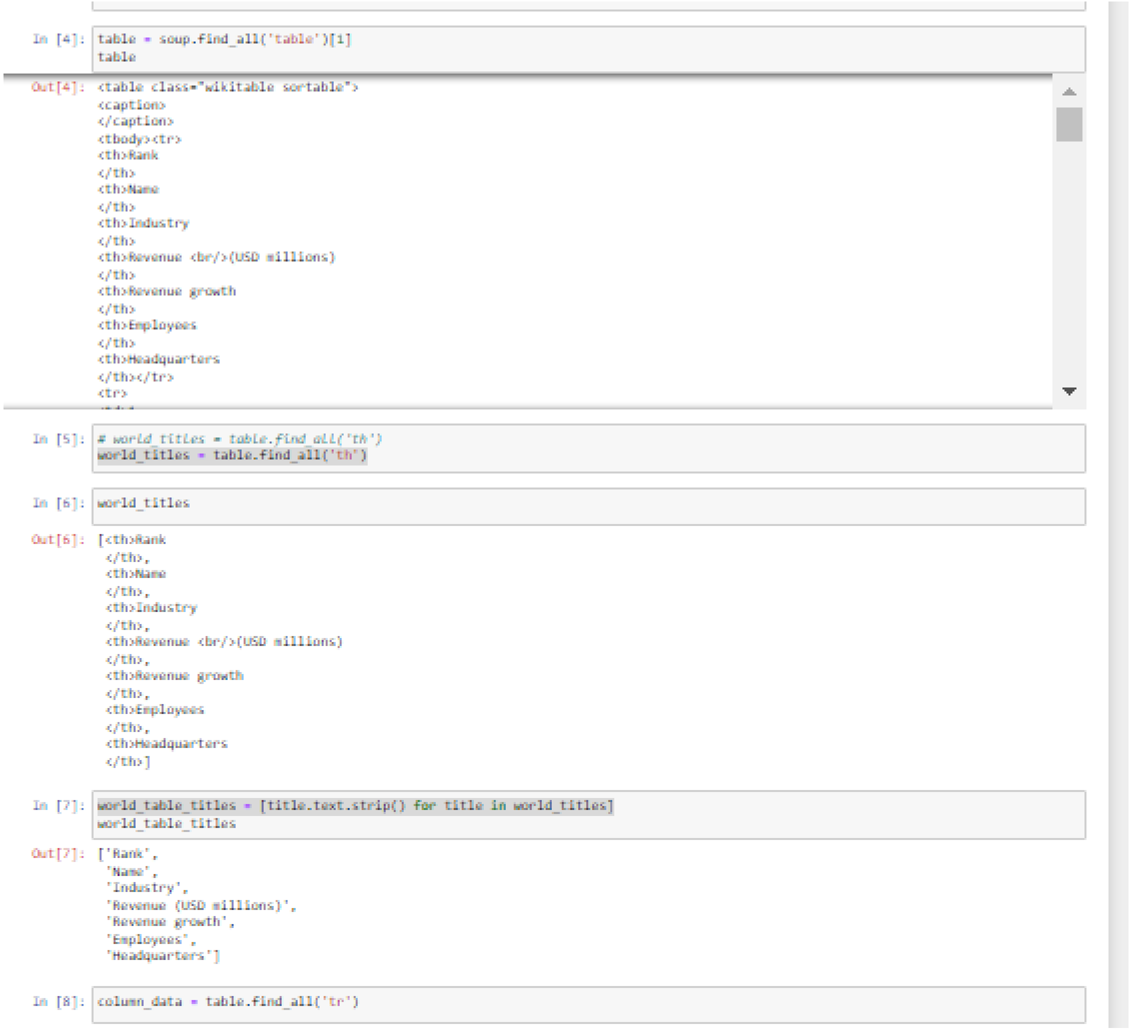


```
In [1]: from bs4 import BeautifulSoup
        import requests
```

```
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 vect
        or-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-fe
        ature-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled vecto
        r-feature-custom-font-size-clientpref-0 vector-feature-client-preferences-disabled vector-feature-client-prefs-pinned-disabl
        ed 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-feature-page-tools-pinned-disabled vector-feature-toc-pinned-cl
        ientpref-1 vector-feature-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-c
        ontent-enabled vector-feature-custom-font-size-clientpref-0 vector-feature-client-preferences-disabled vector-feature-client
        -prefs-pinned-disabled vector-toc-available";var cookie=document.cookie.match(/(?:^|; )enwikimwclientpreferences=([^;]+)/);i
        f(cookie){cookie[1].split('%2C').forEach(function(pref){className=className.replace(new RegExp('(^| )'+pref.replace(/-client
        pref-\w+$|[^\w-]+/g,'')+'-clientpref-\\w+( |$)'),'$1'+pref+'$2');});}document.documentElement.className=className;}());RLCON
        F={"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wg
        MonthNames":["",
        "January","February","March","April","May","June","July","August","September","October","November","December"],"wgRequestI
```



```
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 [6]: world_titles
```

```
Out[6]: [<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>]
```

```
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 [18]:
```
df.set_index('Rank')
```

Out[18]:

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

100 rows × 6 columns

In [38]:
```
print(df.dtypes)
```

```
Rank                      object
Name                      object
Industry                  object
Revenue (USD millions)    object
Revenue growth            object
Employees                 object
Headquarters              object
dtype: object
```

In [55]:
```
df3 = df.copy()
```

In [110]:
```
df3['Employees'] = df3['Employees'].replace(value = '', to_replace = '[^a-zA-Z0-9]', regex = True)
```

In [111]:
```
df3['Revenue (USD millions)'] = df3['Revenue (USD millions)'].replace(value = '', to_replace = '[^a-zA-Z0-9]', regex = True)
```

In [112]:
```
# df3['Revenue growth'] = df3['Revenue growth'].replace(value = '', to_replace = '[^a-zA-Z0-9]', regex = True)
```

In [113]:
```
df3['Employees'] = df3['Employees'].astype(int)
```

In [114]:
```
df3['Revenue (USD millions)'] = df3['Revenue (USD millions)'].astype(int)
```

In [115]:
```
# df3['Revenue growth'] = df3['Revenue growth'].astype(int)
```

```
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')
```

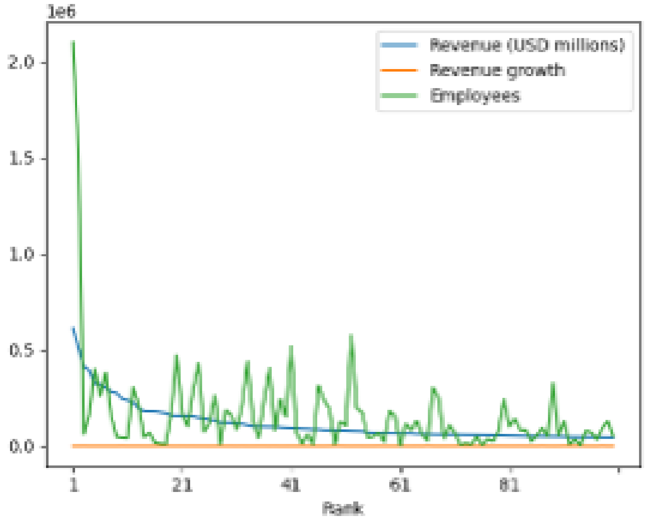
```
In [122]: df.head()
```

Out[122]:

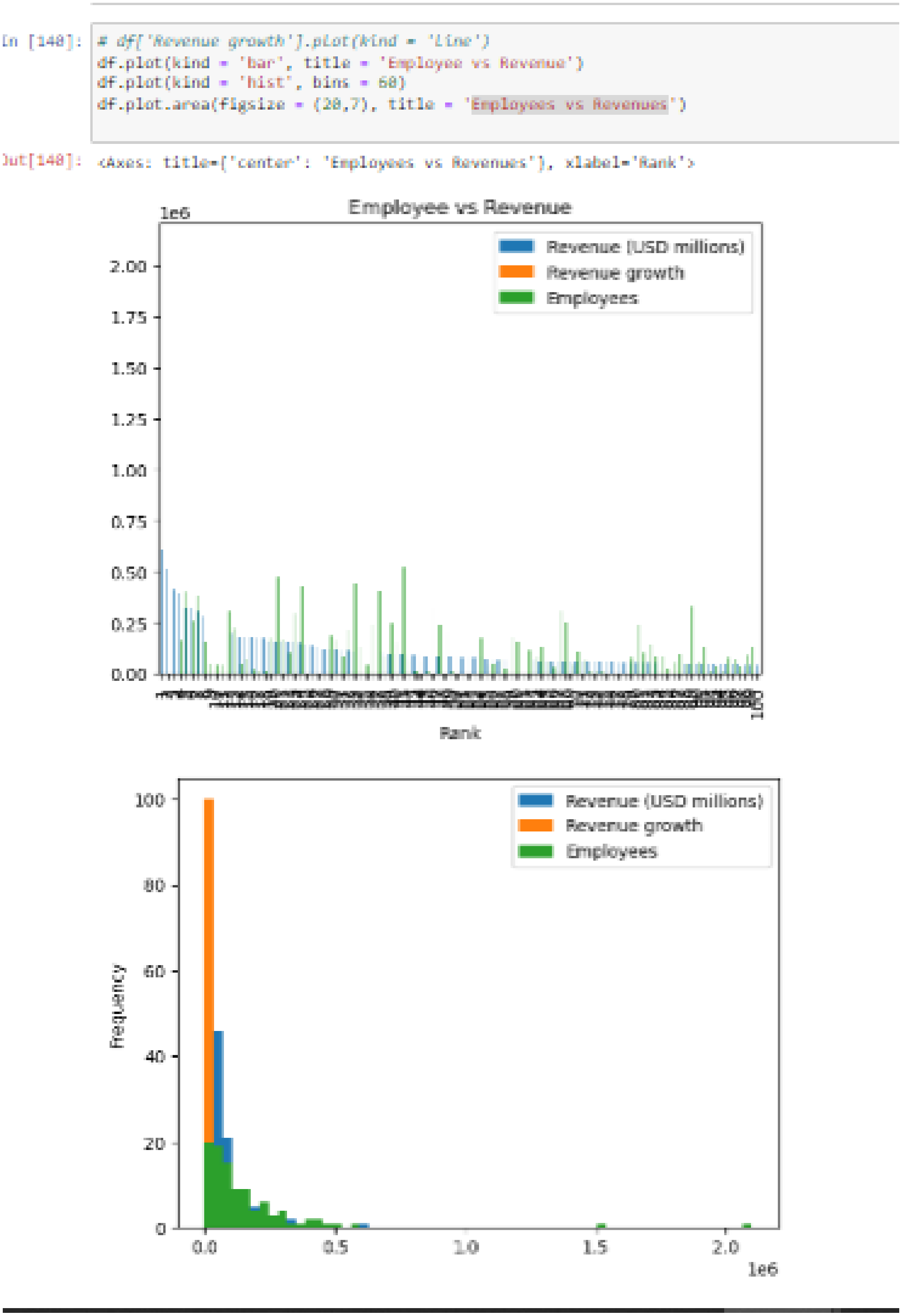
	Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters
Rank						
1	Walmart	Retail	611289	67	2100000	Bentonville, Arkansas
2	Amazon	Retail and cloud computing	513983	94	1540000	Seattle, Washington
3	ExxonMobil	Petroleum industry	413680	448	62000	Spring, Texas
4	Apple	Electronics industry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

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

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

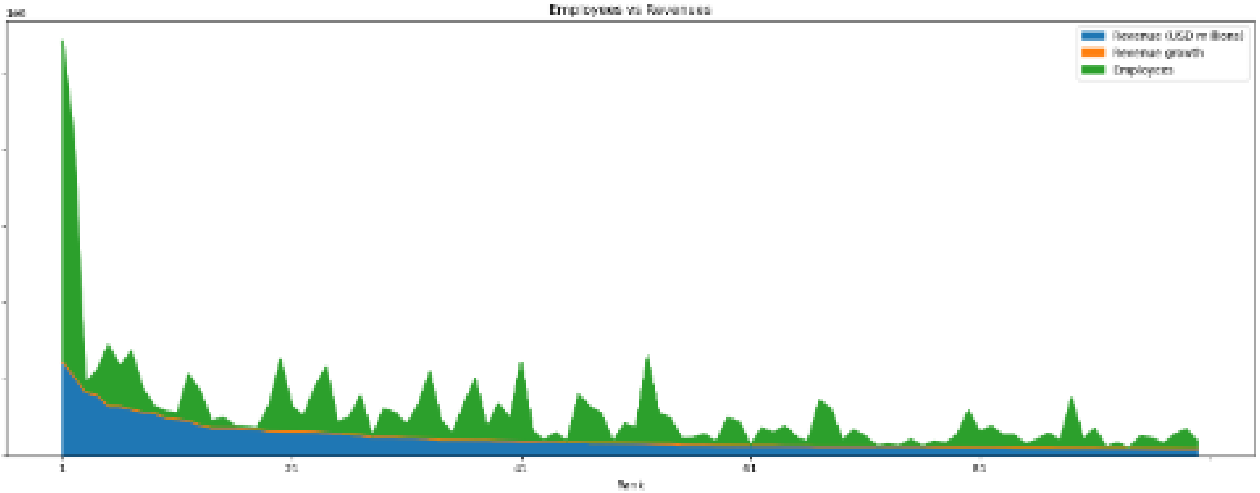


```
In [ ]:
```

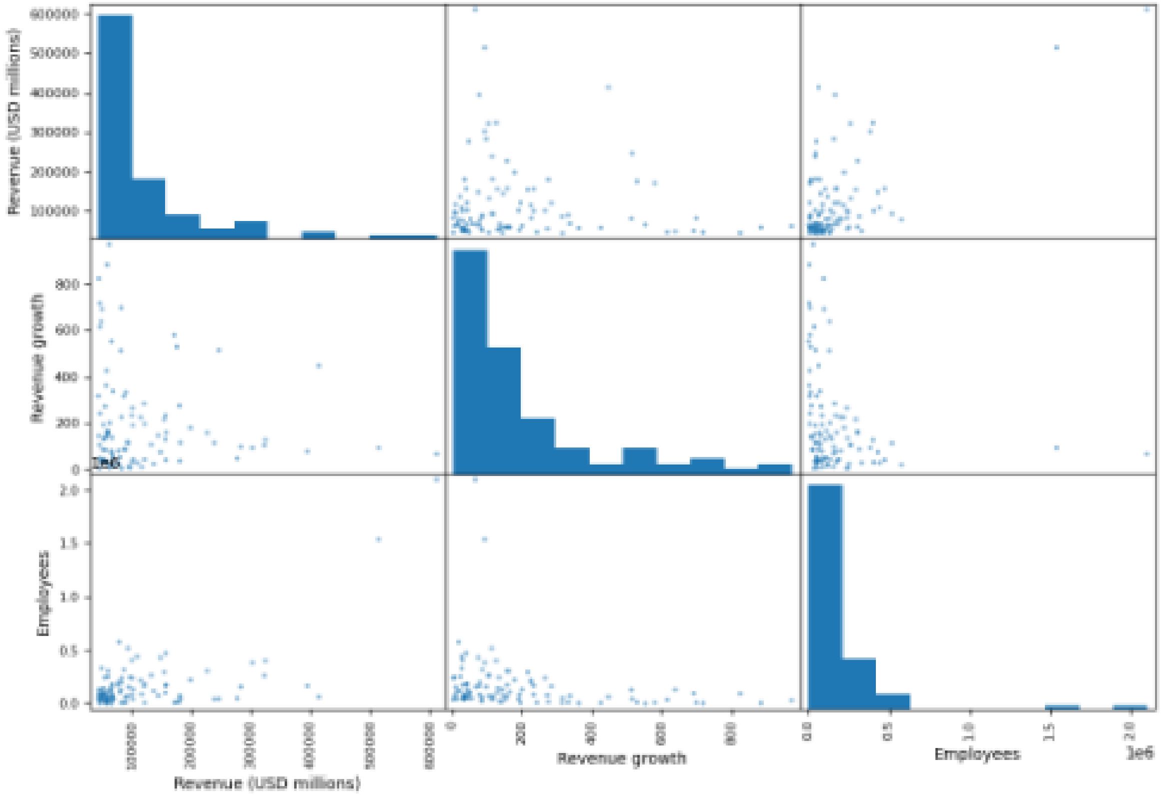


```
In [148]: # df['Revenue growth'].plot(kind = 'Line')
          df.plot(kind = 'bar', title = 'Employee vs Revenue')
          df.plot(kind = 'hist', bins = 60)
          df.plot.area(figsize = (20,7), title = 'Employees vs Revenues')

Out[148]: <Axes: title={'center': 'Employees vs Revenues'}, xlabel='Rank'>
```



```
[<Axes: xlabel='Revenue (USD millions)', ylabel='Employees'>,
 <Axes: xlabel='Revenue growth', ylabel='Employees'>,
 <Axes: xlabel='Employees', ylabel='Employees'>]], dtype=object)
```


Revenue (USD millions)
Revenue growth
1e6
Employees
Revenue (USD millions)
Revenue growth
1e6
Employees