In [2]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:

```python
df=pd.read_csv('Fortune_1000.csv')
df
```

Out[3]:

	company	rank	revenue	profit	num. of employees	sector	city	state	ceo_founder	ceo_woman	CEO	Website	Ticker	Market Cap
0	Walmart	1	572754.0	13673.0	2300000.0	Retailing	Bentonville	AR	no	no	C. Douglas McMillon	https://www.stock.walmart.com	WMT	352037
1	Amazon	2	469822.0	33364.0	1608000.0	Retailing	Seattle	WA	no	no	Andrew R. Jassy	www.amazon.com	AMZN	1202717
2	Apple	3	365817.0	94680.0	154000.0	Technology	Cupertino	CA	no	no	Timothy D. Cook	www.apple.com	AAPL	2443962
3	CVS Health	4	292111.0	7910.0	258000.0	Health Care	Woonsocket	RI	no	yes	Karen Lynch	https://www.cvshealth.com	CVS	125204
4	UnitedHealth Group	5	287597.0	17285.0	350000.0	Health Care	Minnetonka	MN	no	no	Andrew P. Witty	www.unitedhealthgroup.com	UNH	500468
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	Vizio Holding	996	2124.0	-39.4	800.0	Industrials	Irvine	CA	yes	no	William W. Wang	https://www.vizio.com	VZIO	1705.1
996	1-800-Flowers.com	997	2122.2	118.7	4800.0	Retailing	Jericho	NY	no	no	Christopher G. McCann	https://www.1800flowers.com	FLWS	830
997	Cowen	998	2112.8	295.6	1534.0	Financials	New York	NY	no	no	Jeffrey Solomon	https://www.cowen.com	COWN	1078
998	Ashland	999	2111.0	220.0	4100.0	Chemicals	Wilmington	DE	no	no	Guillermo Novo	https://www.ashland.com	ASH	5601.9
999	DocuSign	1000	2107.2	-70.0	7461.0	Technology	San Francisco	CA	no	no	Allan C. Thygesen	https://www.docusign.com	DOCU	21302.8

1000 rows × 14 columns

Data cleaning

In [5]:

```python
df.shape
```

Out[5]:

```
(1000, 14)
```

In [6]:

```python
df.columns
```

Out[6]:

```
Index(['company', 'rank', 'revenue', 'profit', 'num. of employees', 'sector',
       'city', 'state', 'ceo_founder', 'ceo_woman', 'CEO', 'Website', 'Ticker',
       'Market Cap'],
      dtype='object')
```

In [7]:

```python
df.head(10)
```

Out[7]:

	company	rank	revenue	profit	num. of employees	sector	city	state	ceo_founder	ceo_woman	CEO	Website	Ticker	Market Cap
0	Walmart	1	572754.0	13673.0	2300000.0	Retailing	Bentonville	AR	no	no	C. Douglas McMillon	https://www.stock.walmart.com	WMT	352037
1	Amazon	2	469822.0	33364.0	1608000.0	Retailing	Seattle	WA	no	no	Andrew R. Jassy	www.amazon.com	AMZN	1202717
2	Apple	3	365817.0	94680.0	154000.0	Technology	Cupertino	CA	no	no	Timothy D. Cook	www.apple.com	AAPL	2443962
3	CVS Health	4	292111.0	7910.0	258000.0	Health Care	Woonsocket	RI	no	yes	Karen Lynch	https://www.cvshealth.com	CVS	125204
4	UnitedHealth Group	5	287597.0	17285.0	350000.0	Health Care	Minnetonka	MN	no	no	Andrew P. Witty	www.unitedhealthgroup.com	UNH	500468
5	Exxon Mobil	6	285640.0	23040.0	63000.0	Energy	Irving	TX	no	no	Darren W. Woods	www.exxonmobil.com	XOM	371841
6	Berkshire Hathaway	7	276094.0	89795.0	372000.0	Financials	Omaha	NE	no	no	Warren E. Buffett	www.berkshirehathaway.com	BRKA	625468
7	Alphabet	8	257637.0	76033.0	156500.0	Technology	Mountain View	CA	no	no	Sundar Pichai	https://www.abc.xyz	GOOGL	1309359
8	McKesson	9	238228.0	-4539.0	67500.0	Health Care	Irving	TX	no	no	Brian S. Tyler	www.mckesson.com	MCK	47377
9	AmerisourceBergen	10	213988.8	1539.9	40000.0	Health Care	Conshohocken	PA	no	no	Steven H. Collis	www.amerisourcebergen.com	ABC	29972

In [8]:

```python
df.tail(10)
```

Out[8]:

	company	rank	revenue	profit	num. of employees	sector	city	state	ceo_founder	ceo_woman	CEO	Website	Ticker	Market Cap
990	Harsco	991	2147.0	-3.2	12000.0	Business Services	Camp Hill	PA	no	no	F. Nicholas Grasberger III	https://www.harsco.com	HSC	969.7
991	Beazer Homes USA	992	2140.3	122.0	1052.0	Engineering & Construction	Atlanta	GA	no	no	Allan P. Merrill	https://www.beazer.com	BZH	478.8
992	Chemed	993	2139.3	268.6	14137.0	Health Care	Cincinnati	OH	no	no	Kevin J. McNamara	https://www.chemed.com	CHE	7592.5
993	Genesis Energy	994	2125.5	-165.1	1898.0	Energy	Houston	TX	no	no	Grant E. Sims	https://www.genesisenergy.com	GEL	1435.4
994	BWX Technologies	995	2124.1	305.9	6600.0	Aerospace & Defense	Lynchburg	VA	no	no	Rex D. Geveden	https://www.bwxt.com	BWXT	4925.6
995	Vizio Holding	996	2124.0	-39.4	800.0	Industrials	Irvine	CA	yes	no	William W. Wang	https://www.vizio.com	VZIO	1705.1
996	1-800-Flowers.com	997	2122.2	118.7	4800.0	Retailing	Jericho	NY	no	no	Christopher G. McCann	https://www.1800flowers.com	FLWS	830
997	Cowen	998	2112.8	295.6	1534.0	Financials	New York	NY	no	no	Jeffrey Solomon	https://www.cowen.com	COWN	1078
998	Ashland	999	2111.0	220.0	4100.0	Chemicals	Wilmington	DE	no	no	Guillermo Novo	https://www.ashland.com	ASH	5601.9
999	DocuSign	1000	2107.2	-70.0	7461.0	Technology	San Francisco	CA	no	no	Allan C. Thygesen	https://www.docusign.com	DOCU	21302.8

In [9]:

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   company            1000 non-null   object
 1   rank               1000 non-null   int64
 2   revenue            1000 non-null   float64
 3   profit             997 non-null    float64
 4   num. of employees  999 non-null    float64
 5   sector             1000 non-null   object
 6   city               1000 non-null   object
 7   state              1000 non-null   object
 8   ceo_founder        1000 non-null   object
 9   ceo_woman          1000 non-null   object
 10  CEO                1000 non-null   object
 11  Website            1000 non-null   object
 12  Ticker             951 non-null    object
 13  Market Cap         969 non-null    object
dtypes: float64(3), int64(1), object(10)
memory usage: 109.5+ KB
```

In [10]:

```python
df.describe()
```

Out[10]:

	rank	revenue	profit	num. of employees
count	1000.000000	1000.000000	997.000000	9.990000e+02
mean	500.497000	17986.801400	2026.476329	3.578867e+04
std	288.818067	40813.281554	6421.578081	1.046546e+05
min	1.000000	2107.200000	-6520.000000	1.600000e+02
25%	250.750000	3500.750000	195.400000	6.500000e+03
50%	500.500000	6375.500000	572.000000	1.353000e+04
75%	750.250000	14615.475000	1498.300000	2.890000e+04
max	1000.000000	572754.000000	94680.000000	2.300000e+06

In [10]:

```python
df['company'].value_counts
```

Out[10]:

```
<bound method IndexOpsMixin.value_counts of 0                 Walmart
1                  Amazon
2                   Apple
3              CVS Health
4      UnitedHealth Group
              ...
995         Vizio Holding
996     1-800-Flowers.com
997                 Cowen
998               Ashland
999              DocuSign
Name: company, Length: 1000, dtype: object>
```

In [11]:

```python
df['sector'].value_counts
```

Out[11]:

```
<bound method IndexOpsMixin.value_counts of 0        Retailing
1        Retailing
2       Technology
3      Health Care
4      Health Care
          ...
995    Industrials
996      Retailing
997     Financials
998      Chemicals
999     Technology
Name: sector, Length: 1000, dtype: object>
```

In [12]:

```python
df['CEO'].value_counts
```

Out[12]:

```
<bound method IndexOpsMixin.value_counts of 0        C. Douglas McMillon
1            Andrew R. Jassy
2            Timothy D. Cook
3                Karen Lynch
4            Andrew P. Witty
               ...
995          William W. Wang
996    Christopher G. McCann
997          Jeffrey Solomon
998           Guillermo Novo
999        Allan C. Thygesen
Name: CEO, Length: 1000, dtype: object>
```

In [13]:

```python
df['Website'].value_counts
```

Out[13]:

```
<bound method IndexOpsMixin.value_counts of 0      https://www.stock.walmart.com
1                     www.amazon.com
2                      www.apple.com
3          https://www.cvshealth.com
4          www.unitedhealthgroup.com
                   ...
995            https://www.vizio.com
996      https://www.1800flowers.com
997            https://www.cowen.com
998          https://www.ashland.com
999         https://www.docusign.com
Name: Website, Length: 1000, dtype: object>
```

In [13]:

```python
df.sort_values('profit',ascending=False)
df.head(10)
```

Out[13]:

	company	rank	revenue	profit	num. of employees	sector	city	state	ceo_founder	ceo_woman	CEO	Website	Ticker	Market Cap
0	Walmart	1	572754.0	13673.0	2300000.0	Retailing	Bentonville	AR	no	no	C. Douglas McMillon	https://www.stock.walmart.com	WMT	352037
1	Amazon	2	469822.0	33364.0	1608000.0	Retailing	Seattle	WA	no	no	Andrew R. Jassy	www.amazon.com	AMZN	1202717
2	Apple	3	365817.0	94680.0	154000.0	Technology	Cupertino	CA	no	no	Timothy D. Cook	www.apple.com	AAPL	2443962
3	CVS Health	4	292111.0	7910.0	258000.0	Health Care	Woonsocket	RI	no	yes	Karen Lynch	https://www.cvshealth.com	CVS	125204
4	UnitedHealth Group	5	287597.0	17285.0	350000.0	Health Care	Minnetonka	MN	no	no	Andrew P. Witty	www.unitedhealthgroup.com	UNH	500468
5	Exxon Mobil	6	285640.0	23040.0	63000.0	Energy	Irving	TX	no	no	Darren W. Woods	www.exxonmobil.com	XOM	371841
6	Berkshire Hathaway	7	276094.0	89795.0	372000.0	Financials	Omaha	NE	no	no	Warren E. Buffett	www.berkshirehathaway.com	BRKA	625468
7	Alphabet	8	257637.0	76033.0	156500.0	Technology	Mountain View	CA	no	no	Sundar Pichai	https://www.abc.xyz	GOOGL	1309359
8	McKesson	9	238228.0	-4539.0	67500.0	Health Care	Irving	TX	no	no	Brian S. Tyler	www.mckesson.com	MCK	47377
9	AmerisourceBergen	10	213988.8	1539.9	40000.0	Health Care	Conshohocken	PA	no	no	Steven H. Collis	www.amerisourcebergen.com	ABC	29972

In [14]:

```python
df.sort_values('num. of employees',ascending=False)
df.head(10)
```

Out[14]:

	company	rank	revenue	profit	num. of employees	sector	city	state	ceo_founder	ceo_woman	CEO	Website	Ticker	Market Cap
0	Walmart	1	572754.0	13673.0	2300000.0	Retailing	Bentonville	AR	no	no	C. Douglas McMillon	https://www.stock.walmart.com	WMT	352037
1	Amazon	2	469822.0	33364.0	1608000.0	Retailing	Seattle	WA	no	no	Andrew R. Jassy	www.amazon.com	AMZN	1202717
2	Apple	3	365817.0	94680.0	154000.0	Technology	Cupertino	CA	no	no	Timothy D. Cook	www.apple.com	AAPL	2443962
3	CVS Health	4	292111.0	7910.0	258000.0	Health Care	Woonsocket	RI	no	yes	Karen Lynch	https://www.cvshealth.com	CVS	125204
4	UnitedHealth Group	5	287597.0	17285.0	350000.0	Health Care	Minnetonka	MN	no	no	Andrew P. Witty	www.unitedhealthgroup.com	UNH	500468
5	Exxon Mobil	6	285640.0	23040.0	63000.0	Energy	Irving	TX	no	no	Darren W. Woods	www.exxonmobil.com	XOM	371841
6	Berkshire Hathaway	7	276094.0	89795.0	372000.0	Financials	Omaha	NE	no	no	Warren E. Buffett	www.berkshirehathaway.com	BRKA	625468
7	Alphabet	8	257637.0	76033.0	156500.0	Technology	Mountain View	CA	no	no	Sundar Pichai	https://www.abc.xyz	GOOGL	1309359
8	McKesson	9	238228.0	-4539.0	67500.0	Health Care	Irving	TX	no	no	Brian S. Tyler	www.mckesson.com	MCK	47377
9	AmerisourceBergen	10	213988.8	1539.9	40000.0	Health Care	Conshohocken	PA	no	no	Steven H. Collis	www.amerisourcebergen.com	ABC	29972

checking for null values

In [15]:

```python
df.isnull().sum()
```

Out[15]:

```
company               0
rank                  0
revenue               0
profit                3
num. of employees     1
sector                0
city                  0
state                 0
ceo_founder           0
ceo_woman             0
CEO                   0
Website               0
Ticker               49
Market Cap           31
dtype: int64
```

removing null values

In [16]:

```python
df['Market Cap']=df['Market Cap'].fillna(0)
df['profit']=df['profit'].fillna(0)
df['Ticker']=df['Ticker'].fillna(0)
df['num. of employees']=df['num. of employees'].fillna(0)
```

In [14]:

```python
df.isnull().sum()
```

Out[14]:

```
company              0
rank                 0
revenue              0
profit               0
num. of employees    0
sector               0
city                 0
state                0
ceo_founder          0
ceo_woman            0
CEO                  0
Website              0
Ticker               0
Market Cap           0
dtype: int64
```

no of companies in states

In [17]:

```python
state_count= df[['sector','company','state']]
state_count = state_count.groupby(["state"], as_index = False)["company"].count()
state_count.rename(columns = {'company':'company_count'}, inplace=True)
state_count.sort_values(by='company_count',ascending=False,inplace=True)
state_count.head(10)
```

Out[17]:

	state	company_count
3	CA	131
40	TX	97
31	NY	87
13	IL	62
32	OH	54
35	PA	45
8	FL	38
42	VA	34
9	GA	34
18	MA	33

In [18]:

```python
sector_df=df[['company','profit','sector']]
sector_sum_df=sector_df.groupby(['sector'], as_index = False).sum('profit').sort_values(
    by=['profit'], ascending = False)

sector_sum_df.rename(columns = {
    'sector':'sector',
    'profit':'sum_profit'}, inplace = True)

sector_sum_df = sector_sum_df.head(5)

sector_sum_df.head(1000)
```

Out[18]:

	sector	sum_profit
6	Financials	556635.2
17	Technology	459503.3
9	Health Care	209442.4
4	Energy	139828.6
16	Retailing	126835.7

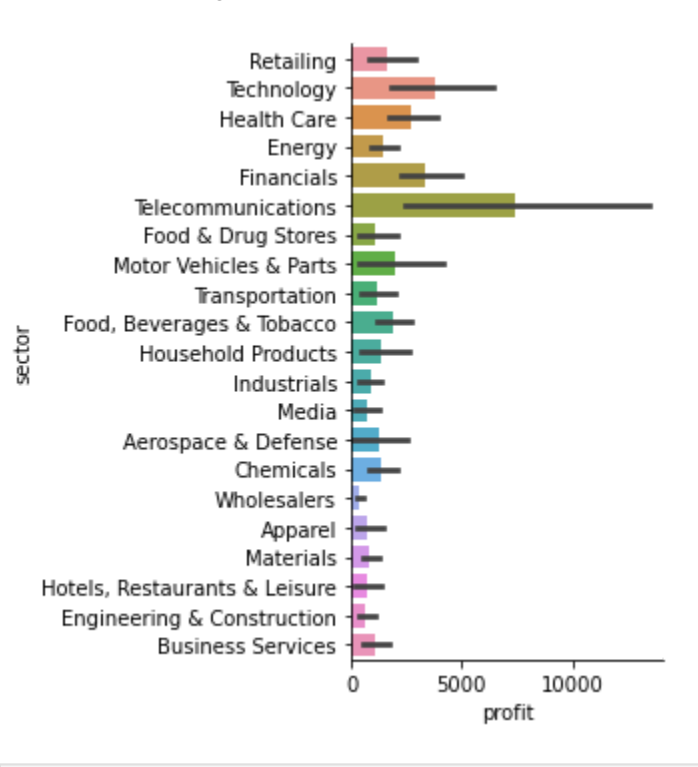
Visualisation

In [19]:

```python
sns.catplot(x="profit",y='sector',data=df,kind="bar")
```

Out[19]:

```
<seaborn.axisgrid.FacetGrid at 0x1b95b78eac0>
```

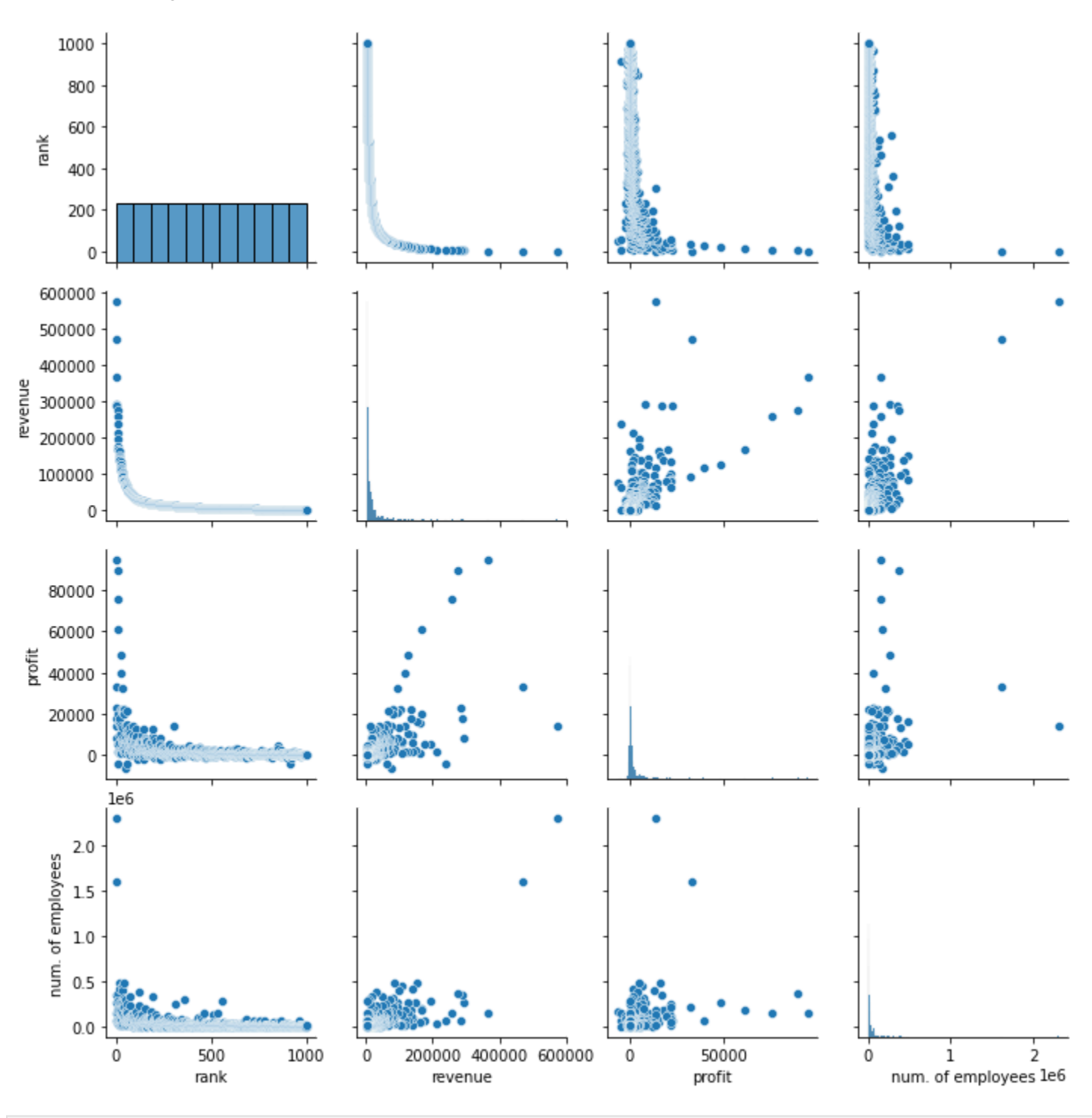


In [20]:

```python
sns.pairplot(df)
```

Out[20]:

```
<seaborn.axisgrid.PairGrid at 0x1b95bbc68b0>
```

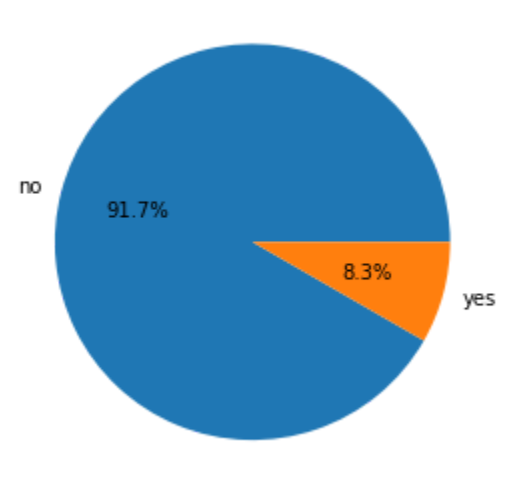


In [14]:

```python
fig = plt.figure(figsize=(10,10))
label1=['no','yes']
ax1 = fig.add_subplot(2, 2, 3)
ax1.pie(df.groupby('ceo_woman')['ceo_woman'].count(),autopct="%1.1f%%",labels=label1)
ax3.set_title('CEO Women')
plt.tight_layout()
plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [14], in <cell line: 5>()
      3 ax1 = fig.add_subplot(2, 2, 3)
      4 ax1.pie(df.groupby('ceo_woman')['ceo_woman'].count(),autopct="%1.1f%%",labels=label1)
----> 5 ax3.set_title('CEO Women')
      6 plt.tight_layout()
      7 plt.show()

NameError: name 'ax3' is not defined
```

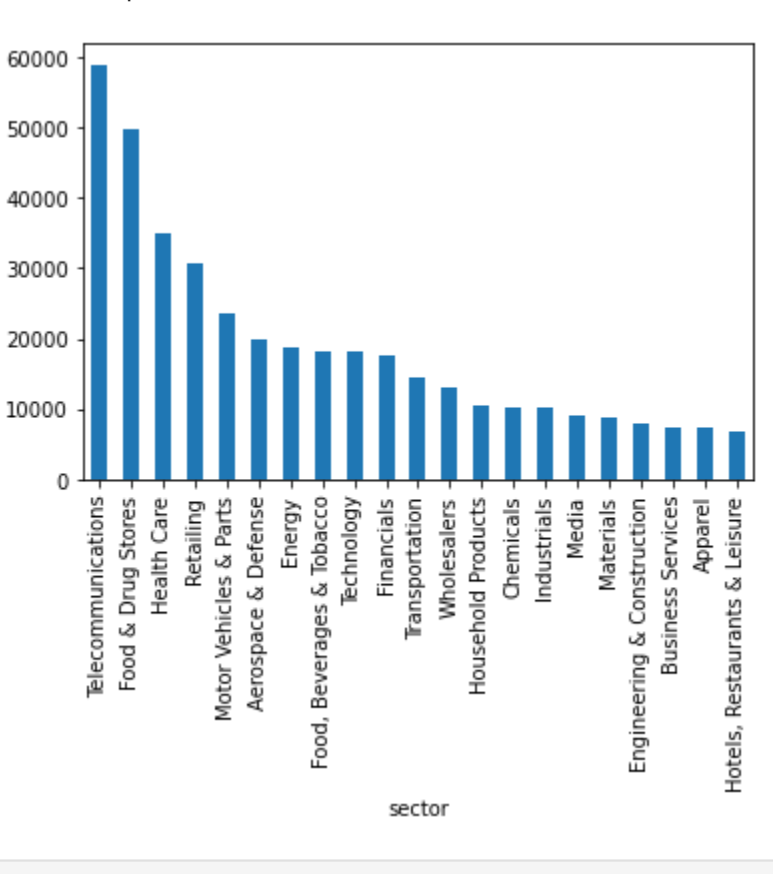


In [37]:

```python
df.groupby('sector')['revenue'].mean().sort_values(ascending=False).plot.bar()
```

Out[37]:

```
<AxesSubplot:xlabel='sector'>
```

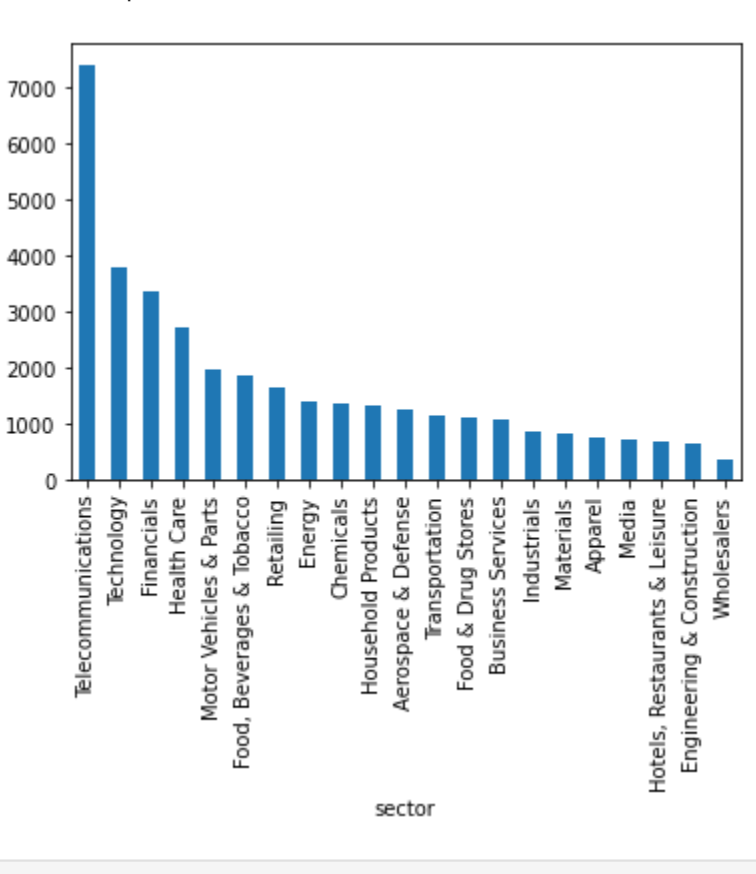


In [23]:

```python
df.groupby('sector')['profit'].mean().sort_values(ascending=False).plot.bar()
```

Out[23]:

```
<AxesSubplot:xlabel='sector'>
```

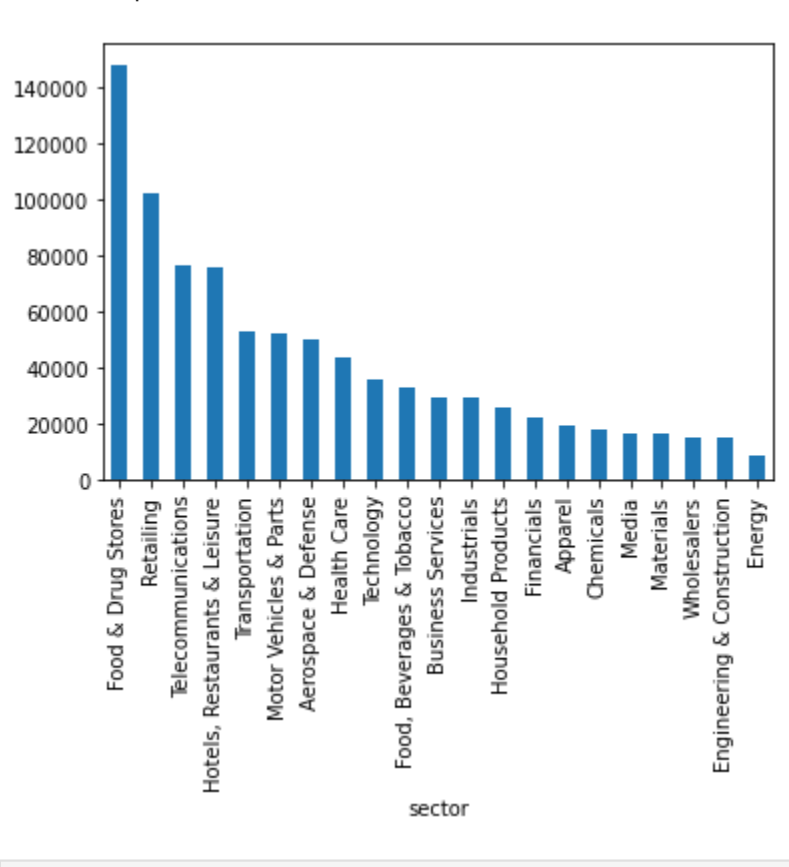


In [35]:

```python
df.groupby('sector')['num. of employees'].mean().sort_values(ascending=False).plot.bar()
```

Out[35]:

```
<AxesSubplot:xlabel='sector'>
```

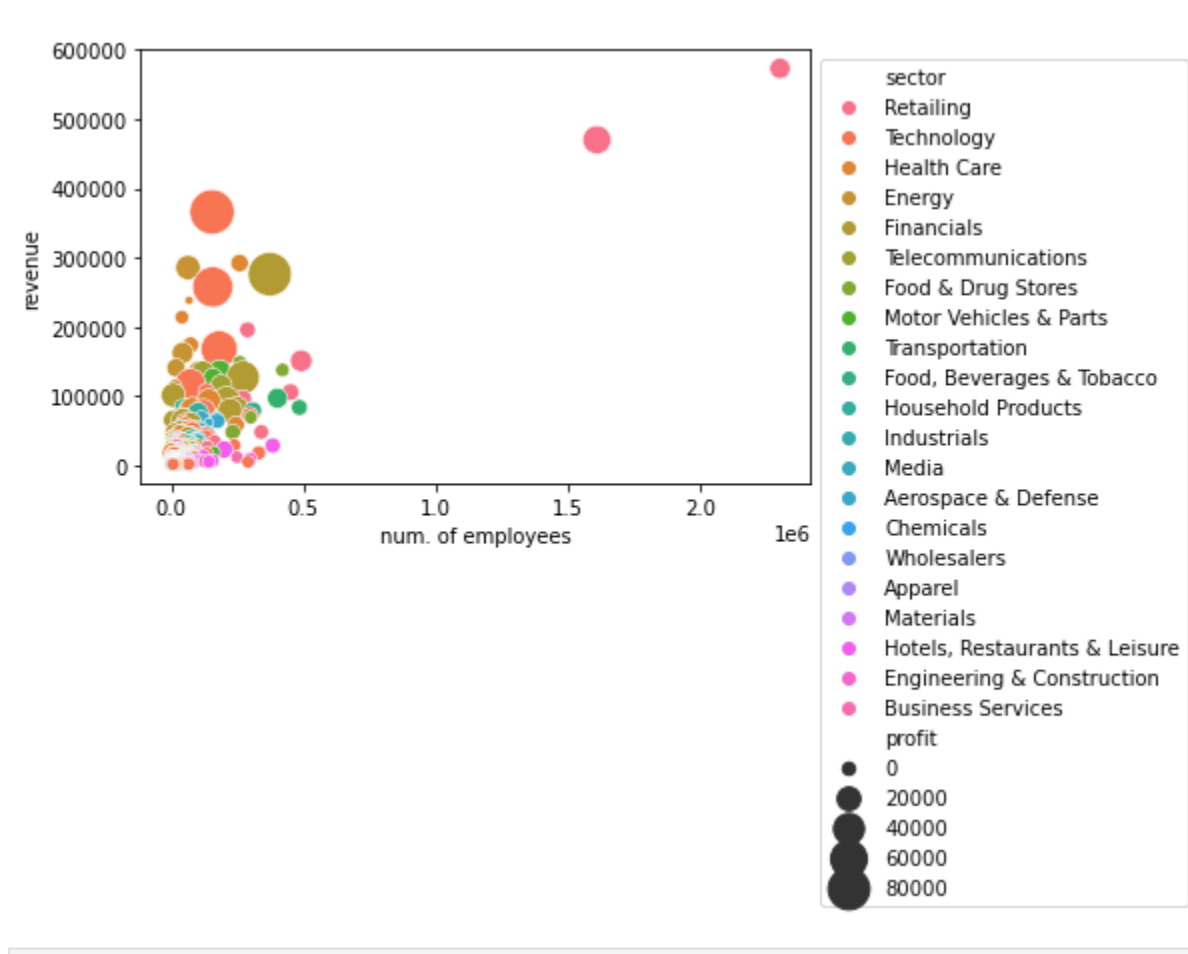


In [34]:

```python
ax=sns.scatterplot(x=df['num. of employees'],y=df['revenue'],hue=df['sector'],size=df['profit'],data=df,sizes=(10,500))
ax.legend(loc='upper left',bbox_to_anchor=(1,1))
```

Out[34]:

```
<matplotlib.legend.Legend at 0x1b963183b50>
```



In [28]:

```
pip install squarify
```

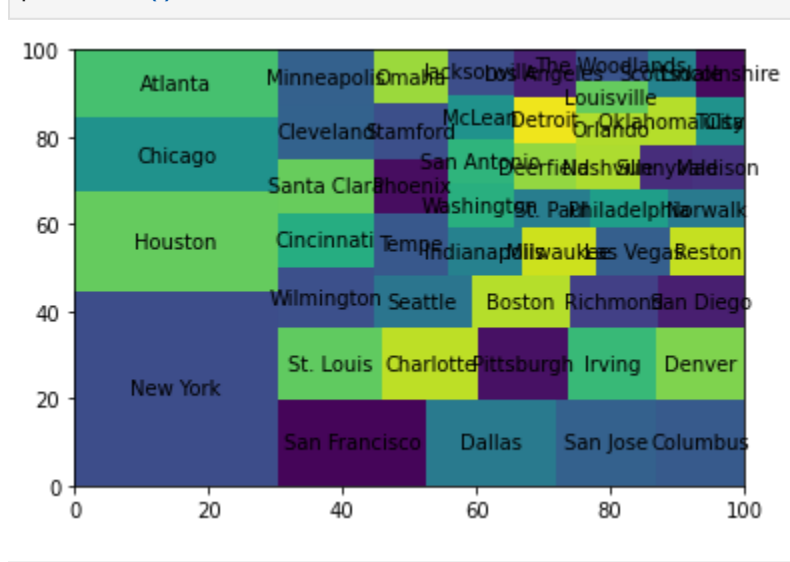
```
Requirement already satisfied: squarify in c:\users\balad\anaconda3\lib\site-packages (0.4.3)
Note: you may need to restart the kernel to use updated packages.
```

In [33]:

```python
import squarify

df_city=pd.DataFrame(df.groupby('city')['city'].count().sort_values(ascending=False).head(50))
df_city=df_city.rename(columns={'city': 'num'})

x = df_city['num']
label = df_city.index
squarify.plot(x, label=label)
plt.axis()
plt.show()
```

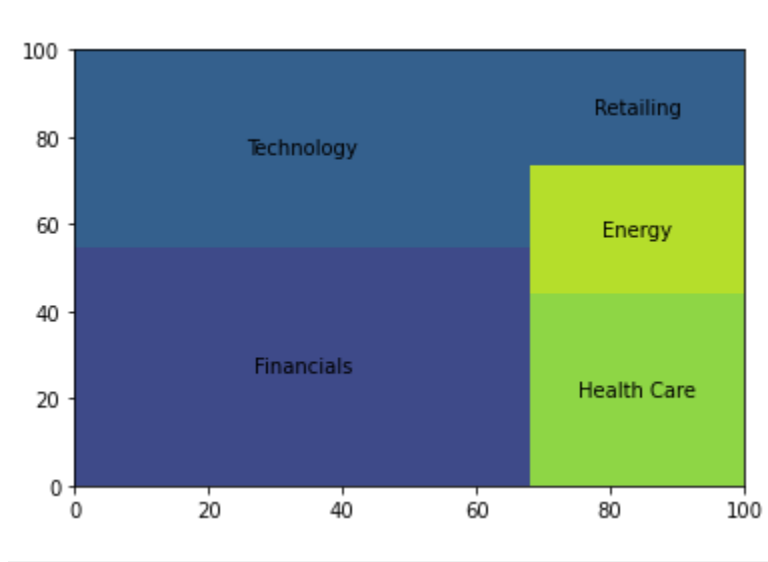


In [32]:

```python
squarify.plot(sizes=sector_sum_df['sum_profit'],
              label=sector_sum_df['sector'],
              text_kwargs={'fontsize':10},
              )
```

Out[32]:

```
<AxesSubplot:>
```



In [31]:

```python
df.corr()
```

Out[31]:

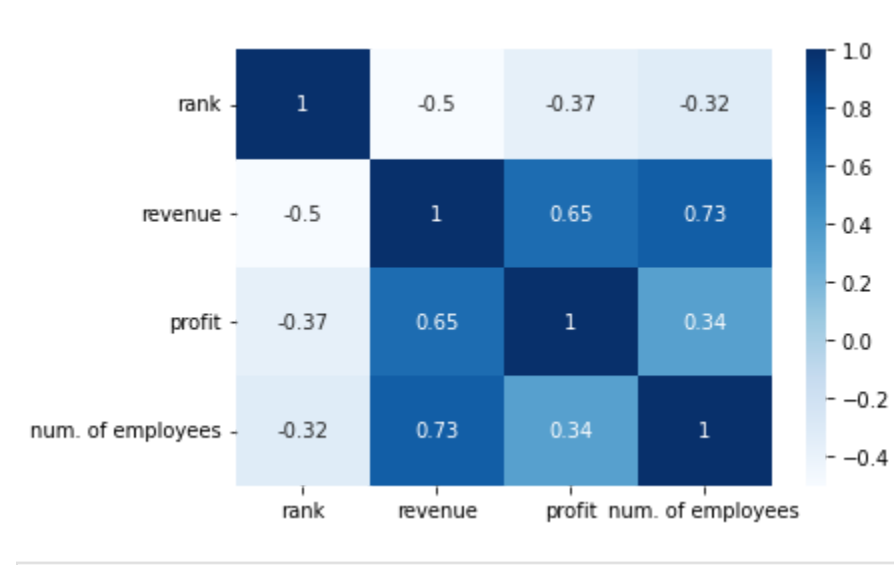
	rank	revenue	profit	num. of employees
rank	1.000000	-0.502607	-0.368651	-0.322330
revenue	-0.502607	1.000000	0.651489	0.732248
profit	-0.368651	0.651489	1.000000	0.337740
num. of employees	-0.322330	0.732248	0.337740	1.000000

In [39]:

```python
sns.heatmap(df.corr(),annot=True,cmap='Blues')
```

Out[39]:

```
<AxesSubplot:>
```



In [ ]: