In [1]:

```
import pandas as pd
import numpy as np
df = pd.read_csv('CS 105 Project.csv')

df = df.set_index(['Year', 'Rank'])

df["Number of Employees"] = df["Number of Employees"].replace({'
\,':''}, regex = True)
df["Number of Employees"] = df["Number of Employees"].replace({'
\,':''}, regex = True)
df["Number of Employees"] = df["Number of Employees"].fillna(0)
df["Number of Employees"] = df["Number of Employees"].astype(flo
at)

df.loc[df['Change in Rank'] == '-', 'Change in Rank'] = 0
df["Change in Rank"] = df["Change in Rank"].fillna(0)
df["Change in Rank"] = df["Change in Rank"].astype(int)

df["Revenues ($millions)"] = df["Revenues ($millions)"].replace(
{'\$':''}, regex = True)
df["Revenues ($millions)"] = df["Revenues ($millions)"].replace(
{'\,':''}, regex = True)
df["Revenues ($millions)"] = df["Revenues ($millions)"].fillna(0
)
df["Revenues ($millions)"] = df["Revenues ($millions)"].astype(f
loat)

df["Revenue Change"] = df["Revenue Change"].replace({'\%':''}, r
egex = True)
df.loc[df['Revenue Change'] == '-', 'Revenue Change'] = 0
df["Revenue Change"] = df["Revenue Change"].fillna(0)
df["Revenue Change"] = df["Revenue Change"].astype(float)

df["Profits ($millions)"] = df["Profits ($millions)"].replace({'
\$':''}, regex = True)
df["Profits ($millions)"] = df["Profits ($millions)"].replace({'
\,':''}, regex = True)
df.loc[df['Profits ($millions)'] == '-', 'Profits ($millions)']
= 0
df["Profits ($millions)"] = df["Profits ($millions)"].fillna(0)
df["Profits ($millions)"] = df["Profits ($millions)"].astype(flo
```

```
at)


df["Profit Change"] = df["Profit Change"].replace({'\%':''}, reg
ex = True)
df.loc[df['Profit Change'] == '-', 'Profit Change'] = 0
df["Profit Change"] = df["Profit Change"].fillna(0)
df["Profit Change"] = df["Profit Change"].astype(float)

df["Assets ($millions)"] = df["Assets ($millions)"].replace({'\$
':''}, regex = True)
df["Assets ($millions)"] = df["Assets ($millions)"].replace({'\,
':''}, regex = True)
df["Assets ($millions)"] = df["Assets ($millions)"].fillna(0)
df["Assets ($millions)"] = df["Assets ($millions)"].astype(float
)

df["Market Value As of 3/29/19 ($m)"] = df["Market Value As of 3
/29/19 ($m)"].replace({'\$':''}, regex = True)
df["Market Value As of 3/29/19 ($m)"] = df["Market Value As of 3
/29/19 ($m)"].replace({'\,':''}, regex = True)
df.loc[df['Market Value As of 3/29/19 ($m)'] == '-', 'Market Val
ue As of 3/29/19 ($m)'] = 0
df["Market Value As of 3/29/19 ($m)"] = df["Market Value As of 3
/29/19 ($m)"].fillna(0)
df["Market Value As of 3/29/19 ($m)"] = df["Market Value As of 3
/29/19 ($m)"].astype(float)

df.head()
#Usman
```

```
/usr/local/lib/python3.6/site-packages/IPython/core/
interactiveshell.py:2848: PerformanceWarning: indexi
ng past lexsort depth may impact performance.
  raw_cell, store_history, silent, shell_futures)
```

Out[1]:

Year	Rank	Company Name	Number of Employees	Change in Rank	Revenues ($millions)	Revenue Change	($m
2019	**1**	Walmart	2200000.0	0	514405.0	2.8	
	2	Exxon Mobil	71000.0	0	290212.0	18.8	2
	3	Apple	132000.0	1	265595.0	15.9	5
	4	Berkshire Hathaway	389000.0	-1	247837.0	2.4	
	5	Amazon.com	647500.0	3	232887.0	30.9	1

Performed data cleaning on the data set.

In [2]:

```
df.describe()
```

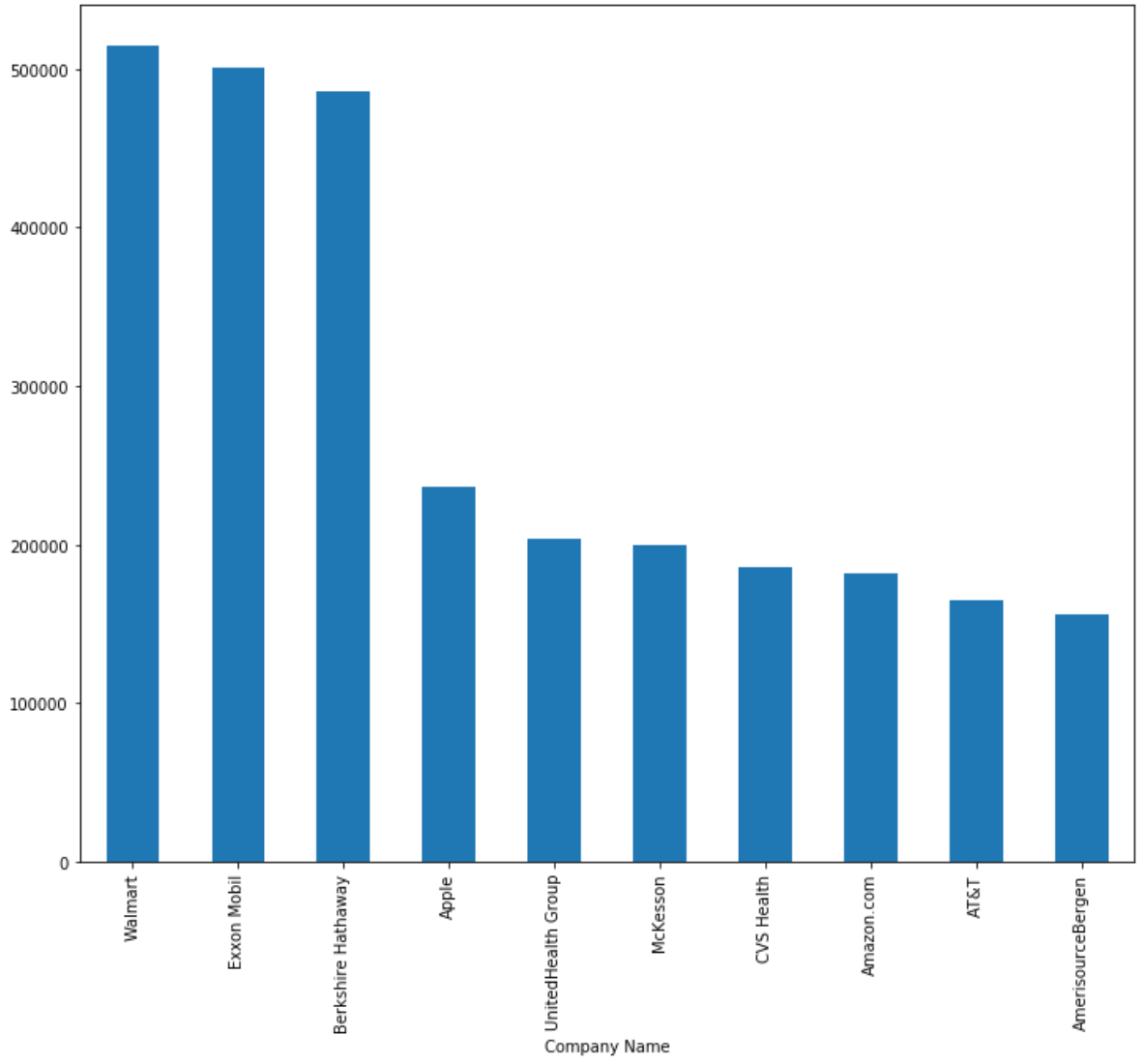
Out[2]:

	Number of Employees	Change in Rank	Revenues ($millions)	Revenue Change	P ($mil
count	1.500000e+03	1500.000000	1500.000000	1500.000000	1500.0(
mean	5.696114e+04	169.955333	25727.784000	8.232400	2020.9-
std	1.238835e+05	172.959504	41020.280072	23.497572	4578.0-
min	1.260000e+02	-131.000000	5145.000000	-57.500000	-22355.0(
25%	1.190000e+04	7.000000	7651.000000	0.000000	288.9[
50%	2.520200e+04	123.500000	12024.000000	5.100000	812.2(
75%	5.809925e+04	312.250000	23387.000000	11.725000	2062.5(
max	2.300000e+06	761.000000	514405.000000	465.300000	59531.0(

This shows the overall statistics of each column in our data.

In [3]:

```
import matplotlib.pyplot as plt
df1 = df[(df['Company Name'] == "Apple") | (df['Company Name'] == "Walmart") | (df['Company Name'] == "Exxon Mobil") | (df['Company Name'] == "Berkshire Hathaway") | (df['Company Name'] == "Amazon.com") | (df['Company Name'] == "UnitedHealth Group") | (df['Company Name'] == "McKesson") | (df['Company Name'] == "CVS Health") | (df['Company Name'] == "AT&T") | (df['Company Name'] == "AmerisourceBergen")]
df1.groupby('Company Name')['Revenues ($millions)'].plot(kind = 'bar', fig=(12,10))
df1.groupby(['Company Name']).mean()['Revenues ($millions)'].sort_values(ascending=False).plot(kind="bar",figsize=(12,10))
plt.show()
#Devang
```



Created a histogram which shows the relationship between the top 10 companies on the Fortune 500 list and their average revenues across the three years.

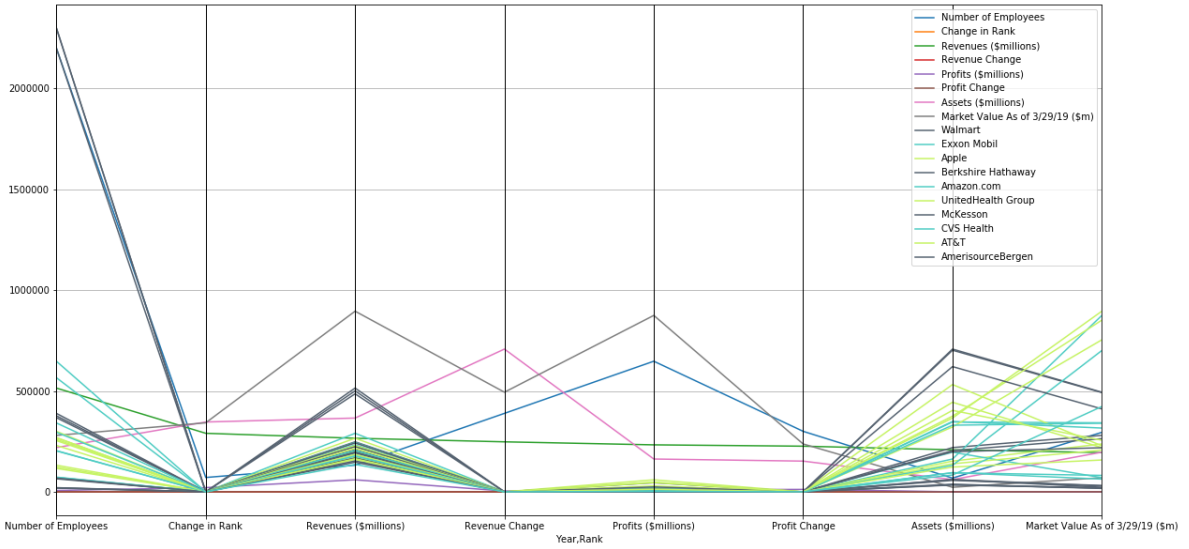
This data appears to be skewed right since the right side of the graph appears to have a longer tail.

In [4]:

```
import matplotlib.pyplot as plt

df1 = df[(df['Company Name'] == "Apple") | (df['Company Name'] == "Walmart") | (df['Company Name'] == "Exxon Mobil") | (df['Company Name'] == "Berkshire Hathaway") | (df['Company Name'] == "Amazon.com") | (df['Company Name'] == "UnitedHealth Group") | (df['Company Name'] == "McKesson") | (df['Company Name'] == "CVS Health") | (df['Company Name'] == "AT&T") | (df['Company Name'] == "AmerisourceBergen")]
df1.plot(figsize=(20,10))
pd.plotting.parallel_coordinates(
        df1, 'Company Name',
        color=('#556270', '#4ECDC4', '#C7F464'))
plt.show()

#Devang
```

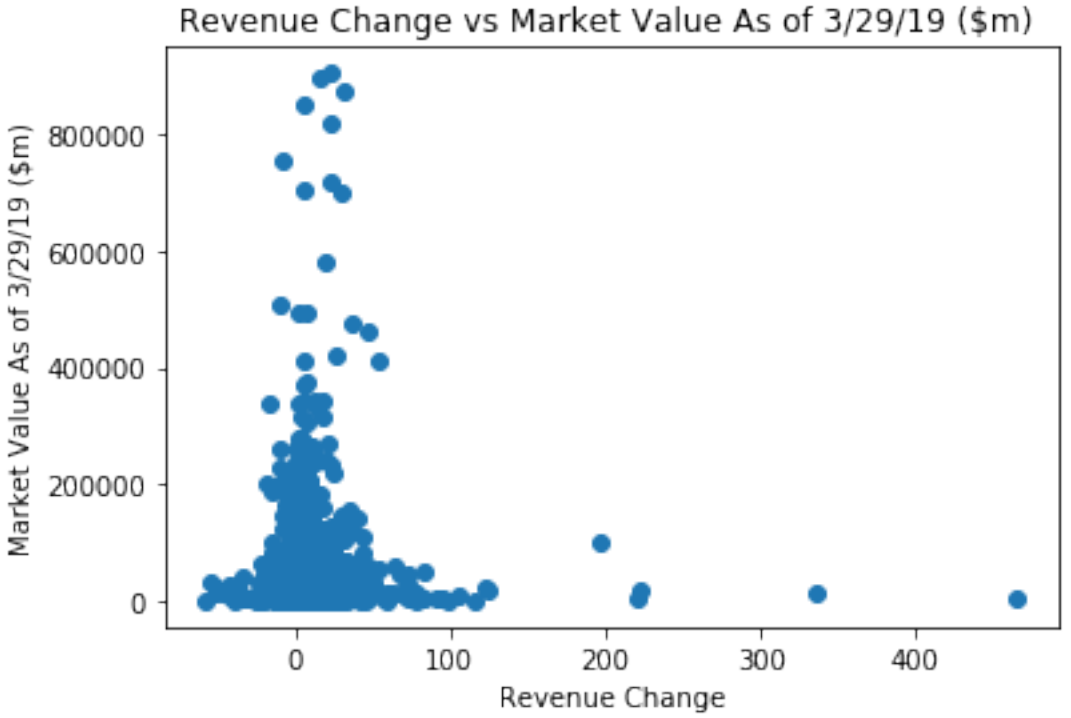


Created a Parallel Coordinates Plot which showed the relationship between the top 10 companies and the other categorical columns.

This graph shows variability and shows the changes among the columns. There seems to be a lot of fluctuation in the top 10 companies.

In [5]:

```
fig=plt.figure()
plt.scatter(df["Revenue Change"], df["Market Value As of 3/29/19 ($m)"])
axis = fig.gca() #get current axis
axis.set_title('Revenue Change vs Market Value As of 3/29/19 ($m) ')
axis.set_xlabel('Revenue Change')
axis.set_ylabel('Market Value As of 3/29/19 ($m)')
fig.canvas.draw()
#Luis
```

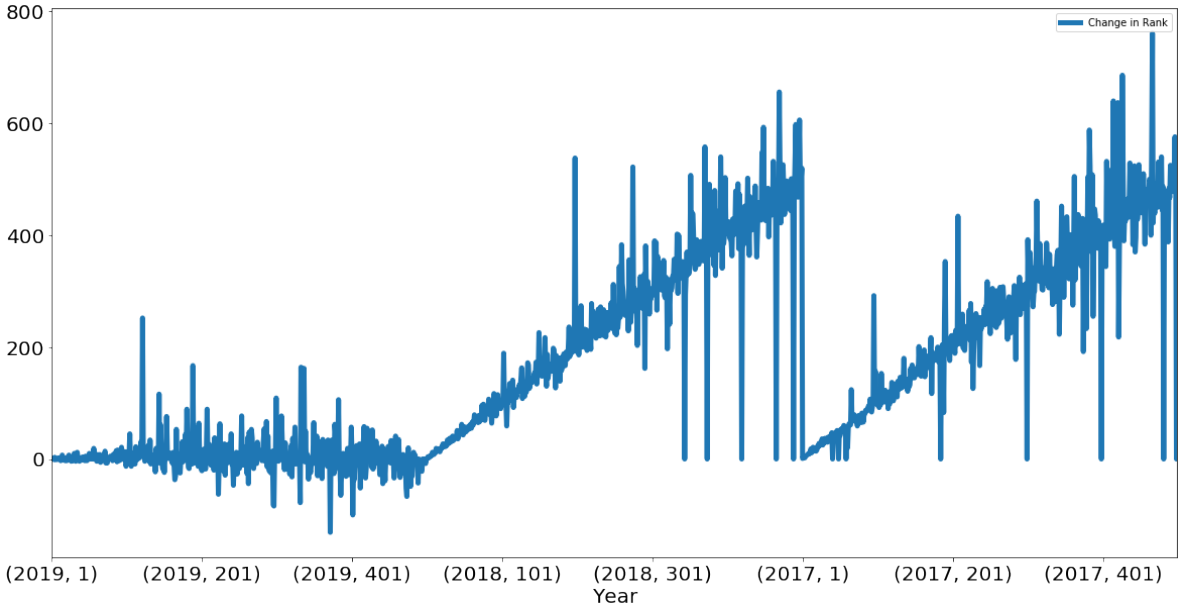


Created a Scatter plot which showed the relationship between Revenue Change and Market Value.

There seems to be no significant correlation between revenue change and market value. The dots seem to clustered together and there are a few outliers.

In [6]:

```
df[['Change in Rank']].plot(figsize=(20,10), linewidth=5, fontsi
ze=20)
plt.xlabel('Year', fontsize=20);
#Luis
```



Created a graph which showed the relationship between Year and Change in Rank.

This seems to be a good representation of how the ranks are being changed on a yearly basis. There seems to be no significant change in the ranks overall.

In [7]:

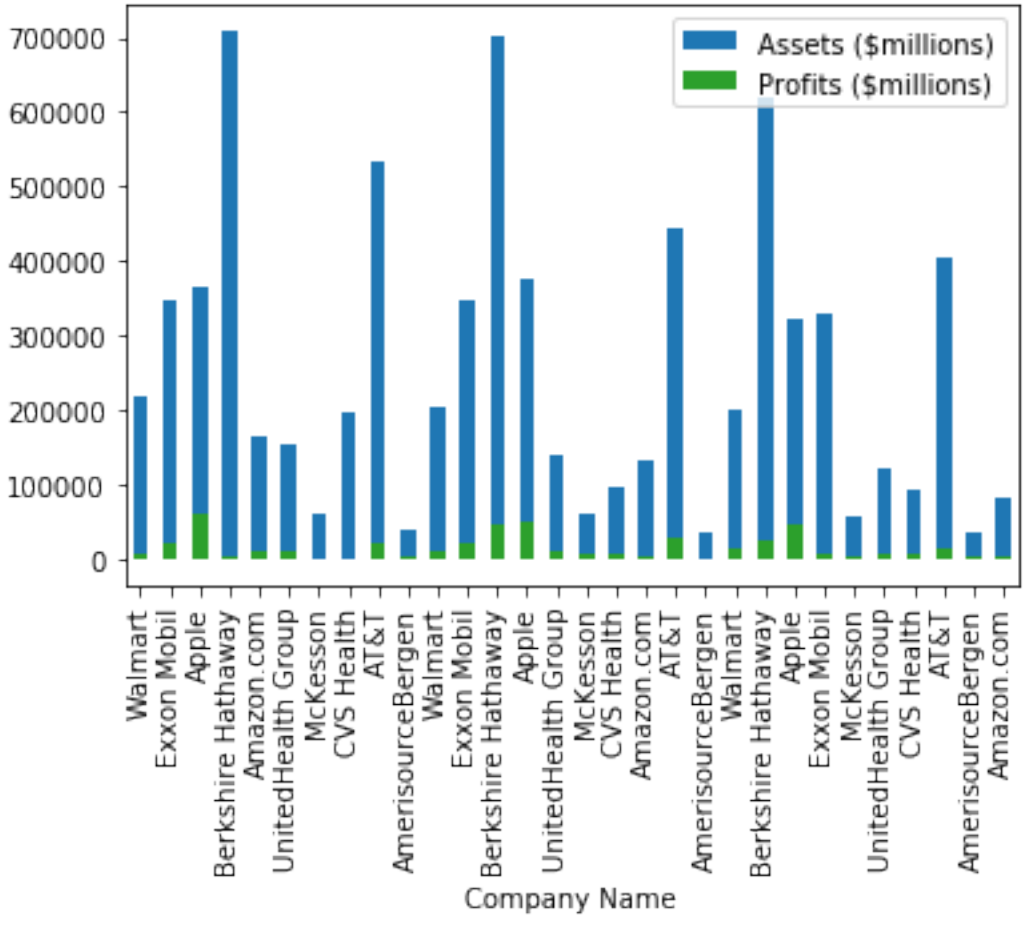
```
df1 = df[(df['Company Name'] == "Apple") | (df['Company Name'] == "Walmart") | (df['Company Name'] == "Exxon Mobil") | (df['Company Name'] == "Berkshire Hathaway") | (df['Company Name'] == "Amazon.com") | (df['Company Name'] == "UnitedHealth Group") | (df['Company Name'] == "McKesson") | (df['Company Name'] == "CVS Health") | (df['Company Name'] == "AT&T") | (df['Company Name'] == "AmerisourceBergen")]

ax = df1.plot(x="Company Name", y="Assets ($millions)", kind="bar")
df1.plot(x="Company Name", y="Profits ($millions)", kind="bar", ax=ax, color="C2")

#Usman
```

Out[7]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb574e3588>
```



Created a graph which showed the relationship between Company Name and Profits and Assets.

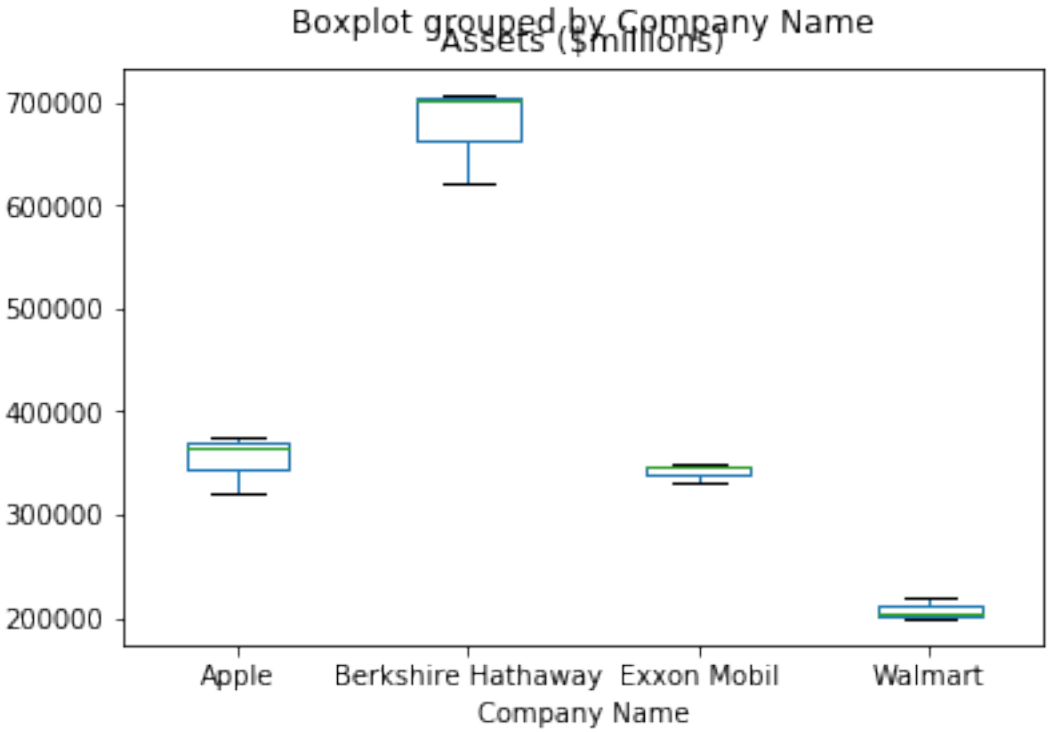
The graph appears to be skewed right.

In [8]:

```
df1 = df[(df['Company Name'] == "Apple") | (df['Company Name'] =
= "Walmart") | (df['Company Name'] == "Exxon Mobil") | (df['Comp
any Name'] == "Berkshire Hathaway")]
df1.boxplot(by ='Company Name', column =['Assets ($millions)'],
grid = False)
#Devang
```

Out[8]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7feb572c
bb00>
```

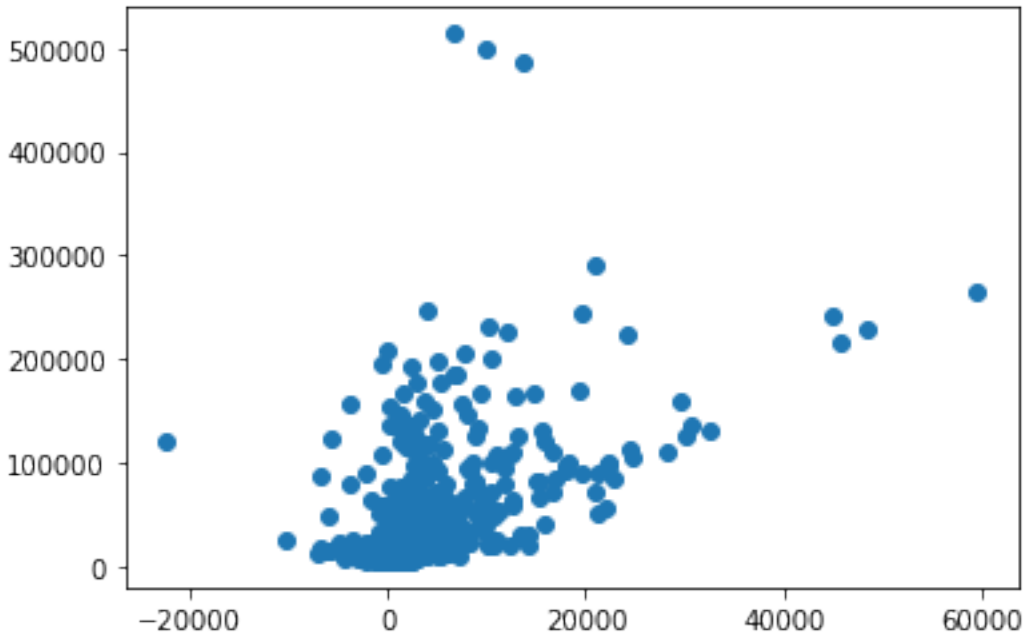


Created a Boxplot which showed the relationship between the top 4 companies and the assets ($millions).

There seems to be a big difference in assets between the four companies. Berkshire Hathaway appears to be leading in terms of assets, even though it is not #1 on the list. Very interesting graph.

In [9]:

```
import matplotlib.pyplot as plt
import pandas
plt.scatter(x=df['Profits ($millions)'], y=df['Revenues ($millio
ns)'])
plt.show()
#Devang
```



Created a Scatter plot which showed the relationship between Profits and Revenues.

There appears to be a great relationship and correlation between revenue and profits. The greater the revenue, the greater the profit. There appears to be a few outliers though.