q2

February 7, 2019

In [25]:
```
#Lisa He
#-*- coding: utf-8 -*-
e_file = open('/Users/Lisa/Documents/BigData/hw1/MS_2007_AR.txt',encoding = "ISO-8859-1
filing =e_file.read()
e_file.close()
```

In [10]:
```
begin_field = filing.find("FINANCIAL HIGHLIGHTS")
begin_field = filing[begin_field:]
end = begin_field.find('Our common stock')
if begin_field != -1:
    cinfo = begin_field[:end]
print(cinfo)
```

```
FINANCIAL HIGHLIGHTS
Ê
```

```
(In millions, except per share data)
```

```
Fiscal Year Ended JuneÊ30
2007
2006
2005
```

2004
2003

Revenue
$51,122
$44,282
$39,788
$36,835
$32,187
Operating income
18,524
16,472
14,561
9,034
9,545
Net income
14,065
12,599
12,254
8,168
7,531
Diluted earnings per share
$ 1.42
$ 1.20
$ 1.12
$ 0.75
$ 0.69
Cash dividends declared per share
$ 0.40
$ 0.35
$ 3.40
$ 0.16
$ 0.08
Cash and short-term investments
23,411
34,161
37,751
60,592
49,048
Total assets
63,171
69,597
70,815

```
94,368
81,732
Long-term obligations
8,320
7,051
5,823
4,574
2,846
StockholdersÕ equity
31,097
40,104
48,115
74,825
64,912
Ê
```

```
In [11]: cinfo
```

```
Out[11]: 'FINANCIAL HIGHLIGHTS \nÊ\n\n\n\n\n\n\n(In millions, except per share data)\n\n\n\n\n\n
```

```
In [15]: #Let's first split the string into lines:
         import re
         cinfo = re.sub("\n+" , "\n", cinfo)
         lines = cinfo.split("\n")
         lines
```

```
Out[15]: ['FINANCIAL HIGHLIGHTS ',
          'Ê',
          '(In millions, except per share data)',
          'Fiscal Year Ended JuneÊ30',
          '2007',
          '2006',
          '2005',
          '2004',
          '2003',
          'Revenue',
          '$51,122',
          '$44,282',
          '$39,788',
          '$36,835',
          '$32,187',
          'Operating income',
          '18,524',
          '16,472',
          '14,561',
          '9,034 ',
          '9,545 ',
```

```
 'Net income',
 '14,065',
 '12,599',
 '12,254',
 '8,168 ',
 '7,531 ',
 'Diluted earnings per share',
 '$    1.42',
 '$    1.20 ',
 '$    1.12 ',
 '$    0.75 ',
 '$    0.69 ',
 'Cash dividends declared per share',
 '$    0.40',
 '$    0.35 ',
 '$    3.40 ',
 '$    0.16 ',
 '$    0.08 ',
 'Cash and short-term investments',
 '23,411',
 '34,161',
 '37,751',
 '60,592',
 '49,048',
 'Total assets',
 '63,171',
 '69,597',
 '70,815',
 '94,368',
 '81,732',
 'Long-term obligations',
 '8,320',
 '7,051 ',
 '5,823 ',
 '4,574 ',
 '2,846 ',
 'StockholdersÕ equity',
 '31,097',
 '40,104',
 '48,115',
 '74,825',
 '64,912',
 'Ê',
 '']
```

In [20]:
```
import re
rev = dict()
oper = dict()
```

```
        total = dict()
        for i in range(0,len(lines)):
            lines[i] = lines[i].strip()
            #Print non empty lines:
            if (lines[i] =='Revenue'):
                for j in range(0,5):
                    var = 2007-j
                    val = lines[i+1+j]
                    rev[var]=val
            if (lines[i] =='Operating income'):
                for j in range(0,5):
                    var = 2007-j
                    val = lines[i+1+j]
                    oper[var]=val
            if (lines[i] =='Total assets'):
                for j in range(0,5):
                    var = 2007-j
                    val = lines[i+1+j]
                    total[var]=val
```

In [24]:
```
print("Revenue:")
print(rev)
print("Operating income:")
print(oper)
print("Total assets:")
print(total)
```

```
Revenue:
{2007: '$51,122', 2006: '$44,282', 2005: '$39,788', 2004: '$36,835', 2003: '$32,187'}
Operating income:
{2007: '18,524', 2006: '16,472', 2005: '14,561', 2004: '9,034', 2003: '9,545'}
Total assets:
{2007: '63,171', 2006: '69,597', 2005: '70,815', 2004: '94,368', 2003: '81,732'}
```