CEO Salary and Performance EDA

Chris Fleisch

September 19, 2016

Introduction

This analysis is motivated by the think tank question:

> Is company performance related to CEO salary?

We will be looking at data provided by the think tank that they have collected from a selection of companies.

Company performance could be measured in a couple different ways. Profit is one way of measuring performance, but a company that doesn't have any profits and increases its market value could also be seen as performing well. Since we only have a single year of market value we won't be able to determine if market value has increased or decreased over time.

We also don't have any previous years of profits. We won't be able to see if a company with negative profits in 1990 had significantly worse previous year's profits and is actually performing well this year even though it still has a negative profit.

We'll have to focus on the positive profits and market value as a leading indicator that the company is performing well in this particular year for this data.

While we're mainly interested in how the CEO's salary is related to company performance we also have the age, college attendance, and number of years at the company and as CEO that could confound our analysis.

This will be an exploratory data analysis focusing on descriptive tools to evaluate relationships between variables. We won't be looking at any causality. For example, we won't be able to say that larger CEO salaries cause companies to perform better.

Setup

We will use the scatterplotMatrix function from the car library to look at all the variables and load ggplot2 library for plotting. The dplyr library will be used for filtering and grouping.

```
library(car)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

This loads the data from the provided Rdata file.

```
load("ceo_w203.RData")
```

Data Structure

We'll look at the structure of the dataset including the number of observations variables and make sure they match what we are expecting.

```
str(CEO)
```

```
## 'data.frame':    184 obs. of  8 variables:
##  $ salary : num  1276 925 2199 369 218 ...
##  $ age    : num  64 56 52 49 57 62 86 59 54 43 ...
##  $ college: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ grad   : num  0 1 1 1 1 0 1 0 0 1 ...
##  $ comten : num  41 26 8 4 33 40 13 35 31 10 ...
##  $ ceoten : num  17 12 8 1 5 6 13 10 4 10 ...
##  $ profits: num  52 67 475 -132 41 -463 11 24 46 48 ...
##  $ mktval : num  1300 2200 6300 1200 421 1400 644 623 812 1100 ...
```

We see 184 rows of observations and 8 variables. This is not a very large sample of data.

All the variables are numbers. We want to convert the college and grad variables to factors since they should only contain a 0 or 1 and are indicating college or grad school attendance.

We convert the college and grad variables to factors by adding them as new variables. This will make the boxplots easier to label later and will help with our summary statistics that we'll look at next.

```
CEO$college.factor <- factor(CEO$college, levels=c(0, 1),
                             labels=c("No college",
                                      "Attended college"))
CEO$grad.factor <- factor(CEO$grad, levels=c(0, 1),
                          labels=c("No grad school",
                                   "Attended grad school"))
str(CEO)
```

```
## 'data.frame':    184 obs. of  10 variables:
##  $ salary        : num  1276 925 2199 369 218 ...
##  $ age           : num  64 56 52 49 57 62 86 59 54 43 ...
##  $ college       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ grad          : num  0 1 1 1 1 0 1 0 0 1 ...
##  $ comten        : num  41 26 8 4 33 40 13 35 31 10 ...
##  $ ceoten        : num  17 12 8 1 5 6 13 10 4 10 ...
##  $ profits       : num  52 67 475 -132 41 -463 11 24 46 48 ...
##  $ mktval        : num  1300 2200 6300 1200 421 1400 644 623 812 1100 ...
##  $ college.factor: Factor w/ 2 levels "No college","Attended college": 2 2 2 2 2 2 2 2 2 2 ...
##  $ grad.factor   : Factor w/ 2 levels "No grad school",..: 1 2 2 2 2 1 2 1 1 2 ...
```

Looking at the structure again we see that we've added college.factor and grad.factor to our data set and set them with appropriate labels.

Next we'll get a summary of all the variables.

```
summary(CEO)
```

```
##      salary            age           college           grad        
##  Min.   : 100.0   Min.   :21.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.: 470.8   1st Qu.:51.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median : 700.5   Median :57.00   Median :1.000   Median :1.0000  
##  Mean   : 856.0   Mean   :55.98   Mean   :0.962   Mean   :0.5489  
##  3rd Qu.:1102.5   3rd Qu.:61.25   3rd Qu.:1.000   3rd Qu.:1.0000  
##  Max.   :5299.0   Max.   :86.00   Max.   :1.000   Max.   :1.0000  
##      comten          ceoten          profits            mktval       
##  Min.   : 2.00   Min.   : 0.000   Min.   :-463.00   Min.   :   -1.0  
##  1st Qu.: 9.00   1st Qu.: 3.000   1st Qu.:  31.75   1st Qu.:  578.2  
##  Median :21.50   Median : 5.500   Median :  57.00   Median : 1200.0  
##  Mean   :21.77   Mean   : 7.755   Mean   : 199.89   Mean   : 3466.1  
##  3rd Qu.:33.00   3rd Qu.:11.000   3rd Qu.: 197.75   3rd Qu.: 3200.0  
##  Max.   :58.00   Max.   :37.000   Max.   :2700.00   Max.   :45400.0  
##           college.factor               grad.factor
##  No college      :  7    No grad school      : 83
##  Attended college:177    Attended grad school:101
##
##
##
##
```

We notice that salary, profits and market value have very large ranges. And salary, profits, and market value all have much larger means than medians. There's likely some large outliers in the data which we'll look at later.

It looks like most of the CEO's attended college while only 7 did not. It's likely that this won't confound with other variables, but we'll take a closer look. The number of CEO's that attended grad school is more evenly split and could confound our analysis. We'll explore that further.

We notice that there's at least one company with a negative profit and one company with a negative market value. We'll need to investigate these further. A company shouldn't have a negative market value.

There's a wide range of ages in this data, but the mean age is around 56. The mean and median years at the company are very close at about 22. And the mean years as CEO is about 8. We'll create some scatter plots to see if age and years at the company relates to profits and market value.

Before we create a scatter plot matrix we should explore some of the interesting variables revealed by the summary. First we'll see how many observations have a negative profit and how many have a negative market value.

```
(nrow(CEO[CEO$profits < 0,]))
```

```
## [1] 15
```

```
(CEO[CEO$profits < 0,])
```

```
##     salary age college grad comten ceoten profits mktval   college.factor
## 168    369  49       1    1      4      1    -132   1200 Attended college
## 114    679  62       1    0     40      6    -463   1400 Attended college
## 4      651  55       1    0     22     22     -54   1000 Attended college
## 52     600  56       1    1     18      7     -40   4000 Attended college
## 42     791  66       1    0     14      8     -60    487 Attended college
## 182    637  45       1    1      3      1      -1     -1 Attended college
## 91     650  55       1    1     28      5    -438    817 Attended college
## 179    677  31       1    1      3      1      -1     -1 Attended college
## 180    173  55       1    1      3      1      -1     -1 Attended college
## 183    877  21       0    1      3      5      -3    303       No college
## 147   1100  65       1    0     18      6    -271    544 Attended college
## 176   2220  63       1    1     18     18     -80    540 Attended college
## 178    379  55       1    1      4      2      -1     -1 Attended college
## 67     630  56       1    0     29      1     -55    420 Attended college
## 181    873  61       1    1      3      1      -1     -1 Attended college
##              grad.factor
## 168 Attended grad school
## 114       No grad school
## 4         No grad school
## 52  Attended grad school
## 42        No grad school
## 182 Attended grad school
## 91  Attended grad school
## 179 Attended grad school
## 180 Attended grad school
## 183 Attended grad school
## 147       No grad school
## 176 Attended grad school
## 178 Attended grad school
## 67        No grad school
## 181 Attended grad school
```

```
(nrow(CEO[CEO$profits == -1,]))
```

```
## [1] 5
```

```
(CEO[CEO$profits == -1,])
```

```
##     salary age college grad comten ceoten profits mktval   college.factor
## 182    637  45       1    1      3      1      -1     -1 Attended college
## 179    677  31       1    1      3      1      -1     -1 Attended college
## 180    173  55       1    1      3      1      -1     -1 Attended college
## 178    379  55       1    1      4      2      -1     -1 Attended college
## 181    873  61       1    1      3      1      -1     -1 Attended college
##              grad.factor
## 182 Attended grad school
## 179 Attended grad school
## 180 Attended grad school
## 178 Attended grad school
## 181 Attended grad school
```

There's 15 observations with negative profits. We'll assume those companies lost money in 1990. Taking a look at the observations with negative values we also see that there's many rows with a -1 for profits. That

seems unusual that there would be 5 with the same -1 value in the profit variable. We then filter for the rows with -1 in profit and we notice that the rows with -1 for profits also have a -1 for market value.

Let's filter for rows with a negative market value.

```
(nrow(CEO[CEO$mktval < 0,]))
```

```
## [1] 5
```

```
(CEO[CEO$mktval < 0,])
```

```
##     salary age college grad comten ceoten profits mktval   college.factor
## 182    637  45       1    1      3      1      -1     -1 Attended college
## 179    677  31       1    1      3      1      -1     -1 Attended college
## 180    173  55       1    1      3      1      -1     -1 Attended college
## 178    379  55       1    1      4      2      -1     -1 Attended college
## 181    873  61       1    1      3      1      -1     -1 Attended college
##              grad.factor
## 182 Attended grad school
## 179 Attended grad school
## 180 Attended grad school
## 178 Attended grad school
## 181 Attended grad school
```

There's five observations that have a market value of -1. They also have a profit of -1. It's likely that these values are unknown and we should code them as NA. This might be a good time to contact the think tank and get clarification on these values. There shouldn't be any companies with a negative market value.

For now, we'll code the -1 values we found in profits and market value as NA and continue the analysis.

```
CEO$profits[CEO$profits == -1] <- NA
CEO$mktval[CEO$mktval == -1] <- NA
summary(CEO)
```

```
##      salary            age           college           grad
##  Min.   : 100.0   Min.   :21.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.: 470.8   1st Qu.:51.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median : 700.5   Median :57.00   Median :1.000   Median :1.0000
##  Mean   : 856.0   Mean   :55.98   Mean   :0.962   Mean   :0.5489
##  3rd Qu.:1102.5   3rd Qu.:61.25   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :5299.0   Max.   :86.00   Max.   :1.000   Max.   :1.0000
##
##      comten          ceoten          profits           mktval
##  Min.   : 2.00   Min.   : 0.000   Min.   :-463.0   Min.   :  200
##  1st Qu.: 9.00   1st Qu.: 3.000   1st Qu.:  34.0   1st Qu.:  616
##  Median :21.50   Median : 5.500   Median :  63.0   Median : 1200
##  Mean   :21.77   Mean   : 7.755   Mean   : 205.5   Mean   : 3563
##  3rd Qu.:33.00   3rd Qu.:11.000   3rd Qu.: 207.0   3rd Qu.: 3350
##  Max.   :58.00   Max.   :37.000   Max.   :2700.0   Max.   :45400
##                                   NA's   :5        NA's   :5
##           college.factor              grad.factor
##  No college      :  7    No grad school      : 83
##  Attended college:177    Attended grad school:101
```

```
##
##
##
##
##
```

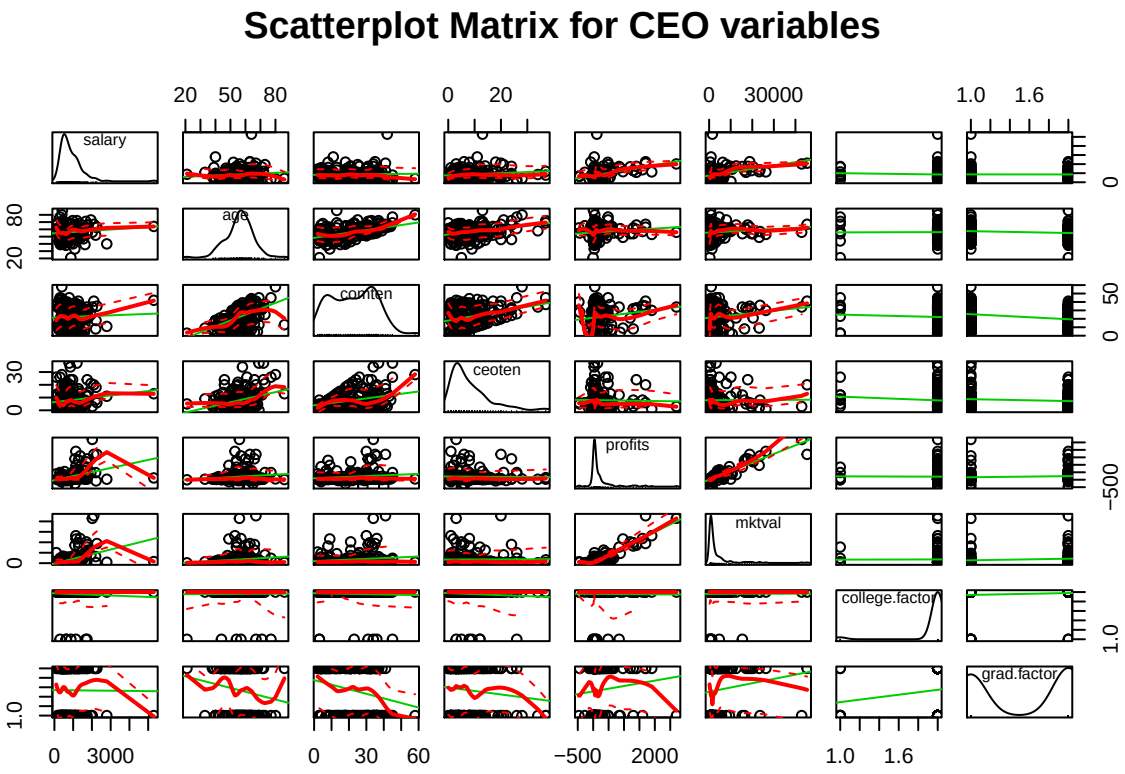
We now have 5 NA's for the values that had a -1 in profits and market value. Our summary shows that our minimum market value is now 200 which sounds more reasonable.

Exploratory Analysis

We have 8 variables and we want to see them all at once to get a quick overview of their relationships. We'll use this as a starting point to dig deeper into the important features of our dataset keeping in mind our question focuses on salary and company performance.

Here's our scatterplot matrix.

```
scatterplotMatrix(~ salary + age + comten + ceoten + profits +
                    mktval + college.factor + grad.factor,
                  data = CEO,
                  main = "Scatterplot Matrix for CEO variables")
```



We notice several relationships for the CEO features that we will use to guide the analysis further:

1. The first thing we notice is that there is a positive relationship between profits and salary. These are the features that we first set out to explore. And there's a similar relationship between market value and salary.

2. There's a strong relationship between age and company years. This is to be expected. The older someone is the more time they have to be working at a company. We see a similar relationship with age and CEO years, but not as strong. Since these relationships are not about market value or profits we won't explore them further. The relationship between age and profits seems to be very minimal and slightly positive. Age and market value have the same slightly positive relationship. We could take a closer look at these later.
3. We see a positive relationship between company years and CEO years, but that's not really what we're interested in. The relationship between company years and profits and market value is only slightly positive.
4. CEO years doesn't appear to have a relationship with profits or market value. The regression lines for these graphs are flat.
5. There's a very positive, almost linear relationship between market value and profits. This is expected since profits usually increase market value.
6. Since most CEOs attended college this variable doesn't seem to have a strong relationship with the other variables.
7. The graduate attendance appears to have a negative relationship with age, company years, CEO years, and a small positive relationship with profits and market value. This variable could confound our analysis and we will want to explore it further.

There's a lot of graphs generated by the scatter plot matrix. It can be hard to tell what's going on unless the graph is larger. We'll generate a correlation matrix to get some numbers to show the strengths of the relationships between our variables.

```
cor(CEO[,1:8], use = "complete.obs")
```

```
##               salary         age      college         grad      comten
## salary   1.000000000  0.11198012 -0.046597338 -0.004924192  0.04041558
## age      0.111980116  1.00000000  0.006984320 -0.144461224  0.49571932
## college -0.046597338  0.00698432  1.000000000  0.101377538 -0.04430620
## grad    -0.004924192 -0.14446122  0.101377538  1.000000000 -0.24046071
## comten   0.040415581  0.49571932 -0.044306202 -0.240460709  1.00000000
## ceoten   0.142754164  0.33033137 -0.078934674 -0.104384252  0.31418312
## profits  0.394231247  0.12433601 -0.010226155  0.091799382  0.15059769
## mktval   0.406615852  0.11708672  0.005587075  0.116750428  0.14313186
##              ceoten     profits      mktval
## salary   0.14275416  0.39423125 0.406615852
## age      0.33033137  0.12433601 0.117086720
## college -0.07893467 -0.01022615 0.005587075
## grad    -0.10438425  0.09179938 0.116750428
## comten   0.31418312  0.15059769 0.143131864
## ceoten   1.00000000 -0.02037427 0.007763170
## profits -0.02037427  1.00000000 0.918373165
## mktval   0.00776317  0.91837317 1.000000000
```

We see some correlation between salary and profits (.39) and salary and market value (.41) which we already suspected as having a strong relationship. And the correlation with profit and market value is very strong (.91) which was displayed in our scatter plot.

Age also shows some correlation with company years (.49) and CEO years (.33) even though we're not really interested in that relationship right now. Age has a small correlation with profits (.12) and market value (.12) as we suspected from the scatter plot matrix.

College doesn't have a strong correlation with profits or market value. Graduate attendance shows a small correlation with profits (.09) and market value (.11).

CEO years doesn't show much correlation between profits and market value. Company years shows a small correlation with profits (.15) and market value (.14).

From the plots we'll want to take a closer look at salary, profits, and market value. We also should see how college and grad school attendance might affect profits and market value. And we can take a look at age and company years.

We'll start by taking a look at our outcome variable profits by summarizing it.

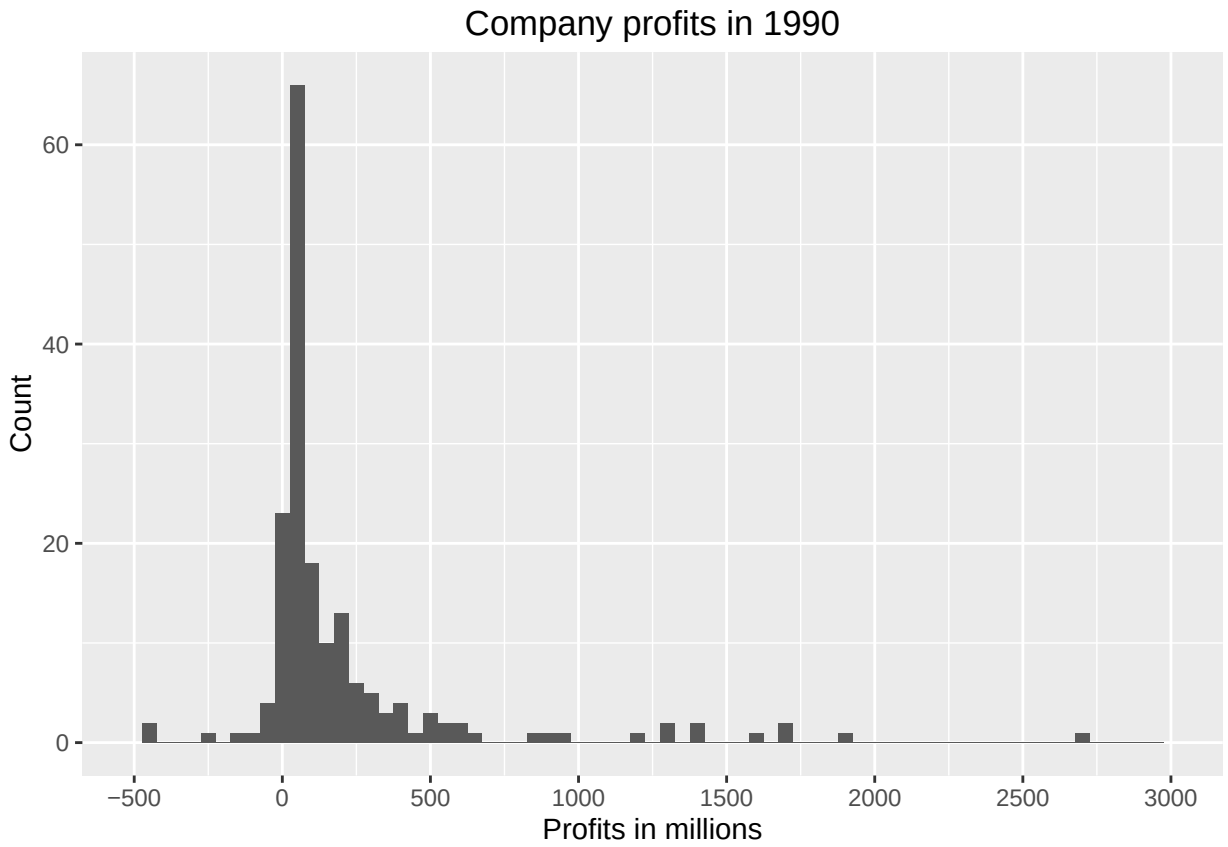
```
summary(CEO$profits)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -463.0    34.0    63.0   205.5   207.0  2700.0       5
```

There's a wide range of profits. There's some negative values and some positive values. The mean is also significantly more than the median which suggests that this data is skewed right.

We'll make a histogram of the profits to get a better understanding of the distribution.

```
qplot(profits, data = na.omit(CEO), binwidth = 50) +
  labs(title = "Company profits in 1990", y = "Count") +
  scale_x_continuous(name = "Profits in millions",
                     limits = c(-500, 3000),
                     breaks = seq(-500, 3000, 500))
```

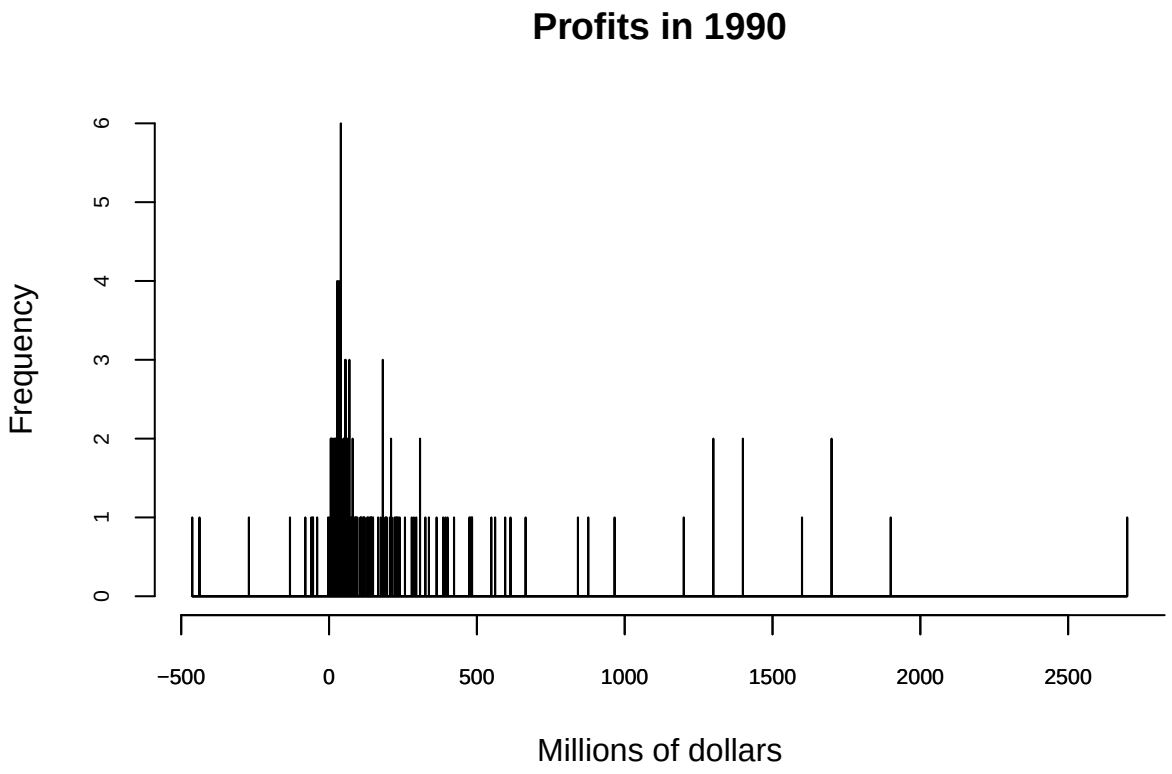


There's a few features to note on the profits histogram:

1. There are some negative values. We suspect that there are a few companies that lost money during the 1990 year.
2. There's a large number of companies with profits in the 0-500 million range. And a large spike around the 0-250 range. We'll need to look at that closer.
3. The data skews right. There's some rather large outliers to the right that are pulling our mean to the right of the median like we saw in the summary.

We should look at another histogram to try and see what's happening at that spike.

```
hist(CEO$profits, breaks = 10000, main = "Profits in 1990",
     xlab = "Millions of dollars", cex.axis = .7)
axis(1, at = seq(-500, 3000, 500), cex.axis = .7)
```



In this histogram we set a large number of breaks so that each value has its own bin and we see several values having the same profit. It's unlikely that each company makes the exact same profit. We should see how many companies are making the same profit.

We'll group the data by profits and get some counts.

```
profits.df = group_by(CEO, profits) %>%
  summarize(counts = n()) %>%
  arrange(desc(counts)) %>%
  as.data.frame()

head(profits.df, 20)
```

```
##    profits counts
## 1       40      6
## 2       NA      5
```

```
## 3       28      4
## 4       34      4
## 5       35      3
## 6       36      3
## 7       55      3
## 8       56      3
## 9       69      3
## 10     182      3
## 11       6      2
## 12       7      2
## 13       8      2
## 14      13      2
## 15      17      2
## 16      21      2
## 17      23      2
## 18      24      2
## 19      30      2
## 20      33      2
```

Grouping by profits we can see that several companies have the exact same values for profits. This seems unusual for a small dataset. We weren't expecting so many overlapping values for a data set with only 184 observations. Since these values are in millions of dollars perhaps the numbers were loose estimations or there was some rounding when the data was entered. We could go back to the think tank to see if we can find out how these numbers where entered.

We should also take a look at some of the other variables of the companies with the same profits to see if any of the other variables are the same.

```
CEO %>% filter(profits == 40)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
## 1    459  59       1    0     33      3      40   1400 Attended college
## 2    650  53       1    1      5      4      40    557 Attended college
## 3    379  51       1    1      9      3      40   1100 Attended college
## 4    393  58       1    1     36      6      40    956 Attended college
## 5   1749  57       1    1     26     11      40  10000 Attended college
## 6    650  69       1    0     37     13      40    817 Attended college
##            grad.factor
## 1       No grad school
## 2 Attended grad school
## 3 Attended grad school
## 4 Attended grad school
## 5 Attended grad school
## 6       No grad school
```

For the 6 companies that made 40 million in profits, most of other variables are different. We see two CEOs made 650. We should probably check the salary variable too.

We'll also take a look at the next highest profit group. These companies made 28 million in profit.

```
CEO %>% filter(profits == 28)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
```

```
## 1    497  44       1    1      8      6      28    387 Attended college
## 2    387  71       1    1     32     13      28    477 Attended college
## 3    129  66       1    1      4      4      28    412 Attended college
## 4    270  43       1    0     15      2      28    713 Attended college
##            grad.factor
## 1 Attended grad school
## 2 Attended grad school
## 3 Attended grad school
## 4       No grad school
```

For companies that made 28 million in profits their other variables are different from each other. It doesn't look like it's a case of the same observations entered multiple times.

We will take a look at the summary of market value.

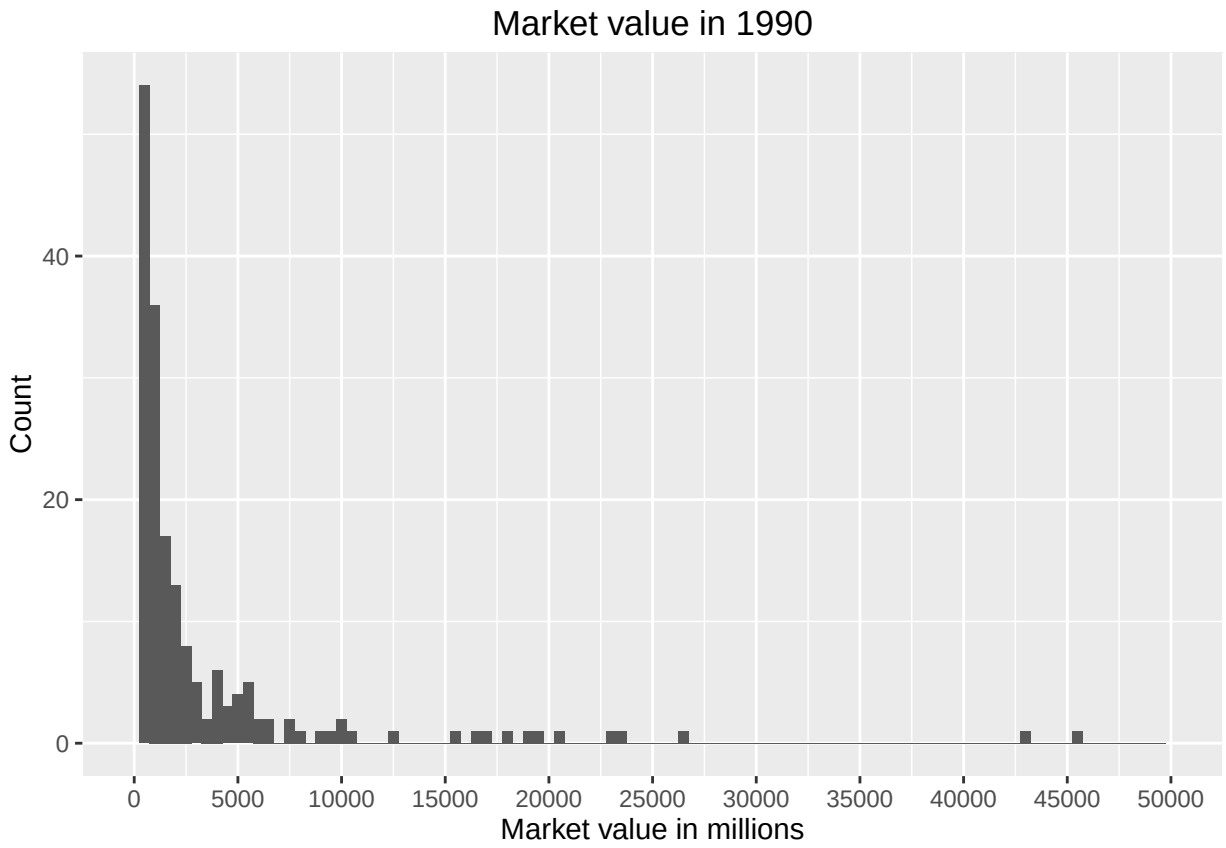
```
summary(CEO$mktval)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     200     616    1200    3563    3350   45400       5
```

The mean is significantly greater than the median. The data is skewed right. There is a large range of values from 200 - 45400. All the values are positive after we coded the negative values as NA.

Let's look at a histogram of the market value.

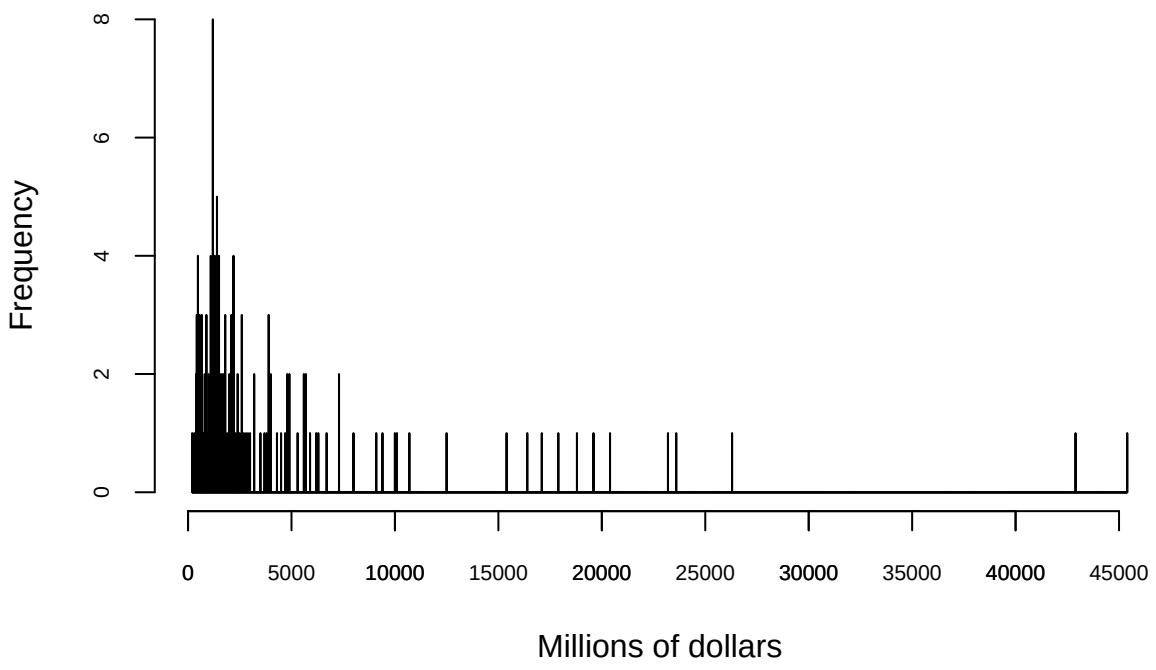
```
qplot(mktval, data = na.omit(CEO), binwidth = 500) +
  labs(title = "Market value in 1990", y = "Count") +
  scale_x_continuous(name = "Market value in millions",
                     limits = c(0, 50000),
                     breaks = seq(0, 50000, 5000))
```



A few features stand out:

1. The data is heavily skewed right with a significant amount of market values less than a billion.
2. We notice that all values are positive after we removed the -1 values.
3. There is another large spike in this histogram. We should make another histogram to explore this further.

```
hist(CEO$mktval, breaks = 10000, main = "Market value in 1990",
     xlab = "Millions of dollars", cex.axis = .7)
axis(1, at = seq(0, 45000, 5000), cex.axis = .7)
```



With a large number of breaks we see that there are some companies that have the same market value.

We'll explore more and group the market values together to try and see how many have the same market value.

```
market_val.df = group_by(CEO, mktval) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  as.data.frame()

head(market_val.df, 20)
```

```
##    mktval count
## 1    1200     8
## 2    1400     5
## 3      NA     5
## 4    1100     4
## 5    1300     4
## 6    1500     4
## 7    2200     4
## 8    1800     3
## 9    2100     3
## 10   2600     3
## 11   3900     3
## 12    420     2
## 13    477     2
## 14    533     2
## 15    686     2
## 16    817     2
## 17    880     2
```

```
## 18   1000     2
## 19   1600     2
## 20   1700     2
```

Here we see a significant amount of companies with the same market value. We could check with the think tank again about these values. It could be another issue of rounding or estimation when the values were entered into the sample. We see 8 companies with the same market value of 1200. With 4 digits there is room for more precision. It could be that the exact value is not interesting for this question and we might only need rough estimates. For other types of questions this might be problematic.

Let's explore some of the rows that have the same market value.

```
filter(CEO, mktval == 1200)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
## 1    369  49       1    1      4      1    -132   1200 Attended college
## 2    622  57       1    0     35      4     143   1200 Attended college
## 3   1119  61       1    0     34      9      71   1200 Attended college
## 4   1101  62       1    1     32      3      96   1200 Attended college
## 5    541  51       1    0     30      4      82   1200 Attended college
## 6    707  46       1    1      6      1      26   1200 Attended college
## 7    377  45       1    0      7      5      57   1200 Attended college
## 8   1675  71       0    0     31     12     115   1200       No college
##            grad.factor
## 1 Attended grad school
## 2       No grad school
## 3       No grad school
## 4 Attended grad school
## 5       No grad school
## 6 Attended grad school
## 7       No grad school
## 8       No grad school
```

```
filter(CEO, mktval == 1400)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
## 1    679  62       1    0     40      6    -463   1400 Attended college
## 2    459  59       1    0     33      3      40   1400 Attended college
## 3    720  49       1    0     12     12      23   1400 Attended college
## 4    348  43       1    1     12     10      79   1400 Attended college
## 5   1041  63       1    1     21     11      91   1400 Attended college
##            grad.factor
## 1       No grad school
## 2       No grad school
## 3       No grad school
## 4 Attended grad school
## 5 Attended grad school
```

When looking at some of companies that have the same market values we don't see many other values that are the same across all the other variables. We'll assume that these values are not duplicates. It does seem odd that these values are the same when the market value obtained from a source like the stock market would have more exact figures.

We'll next look at salary information for the CEOs. Here's a summary of salary.

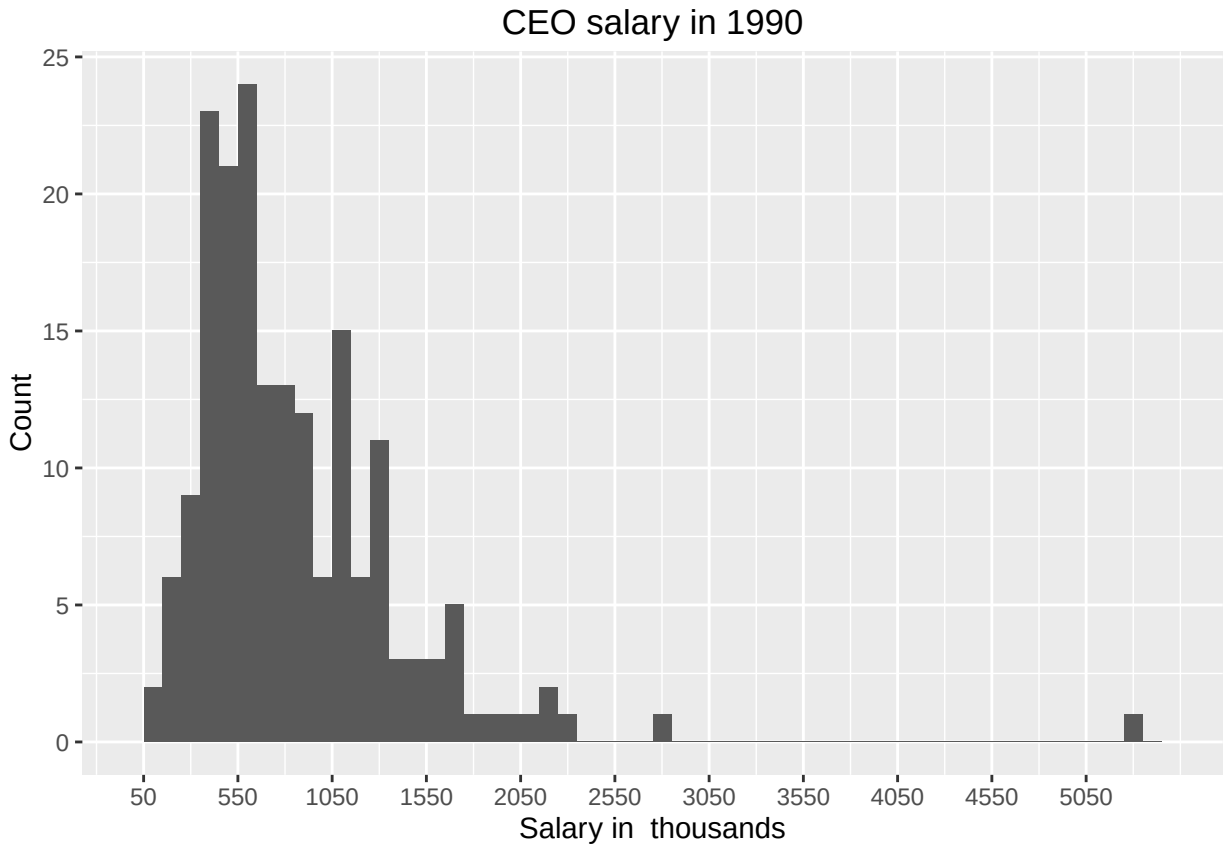
```
summary(CEO$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   470.8   700.5   856.0  1102.0  5299.0
```

The salary summary shows us that the mean is larger than the median suggesting that this data skews right. We also notice that 50% of the values are in 470 - 1102 range and that there's one max value over 5 million. The CEOs in this sample all seem to be getting by OK and since they are running companies we expect them to make large salaries.

We'll make a histogram of the salary.

```
qplot(salary, data = CEO, binwidth = 100) +
  labs(title = "CEO salary in 1990", y = "Count") +
  scale_x_continuous(name = "Salary in  thousands",
                     limits = c(0, 5500),
                     breaks = seq(50, 5500, 500))
```



Features to note in the salary histogram:

1. This histogram shows all values are positive (as expected for a salary).
2. The graph skews towards the right. It could be interesting to see if this reimbursement is appropriate for the company's performance based on other CEO salaries and performance. We won't explore that here.
3. Most of the values are less than 2 million.

4. We see a couple spikes in the data. There's a couple large ones in the 400-800 range. And then a few lower spikes as the graph moves towards the right. There could be a common CEO level of compensation based on company size that might account for the spikes.
5. From this histogram we can see that the salary ramps up pretty quickly at about 400 and then is right skew. This might suggest that most CEOs are starting at about 400,000 for this position and CEO's making less may want to use this to negotiate better compensation. CEO's making over 1.5 million should be considered pretty well compensated for this particular sample.

We will group the salaries together to see how many are making the same amount.

```
salaries.df = group_by(CEO, salary) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  as.data.frame()

head(salaries.df, 20)
```

```
##    salary count
## 1     650     3
## 2     358     2
## 3     379     2
## 4     474     2
## 5     537     2
## 6     559     2
## 7     600     2
## 8     609     2
## 9     637     2
## 10    693     2
## 11    834     2
## 12    873     2
## 13   1142     2
## 14    100     1
## 15    129     1
## 16    173     1
## 17    174     1
## 18    185     1
## 19    218     1
## 20    245     1
```

There are number of salaries that are the exact same, but the highest count is 3 for 650. We'll take a closer look at the top 3.

```
filter(CEO, salary == 650)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
## 1    650  53       1    1      5      4      40    557 Attended college
## 2    650  55       1    1     28      5    -438    817 Attended college
## 3    650  69       1    0     37     13      40    817 Attended college
##            grad.factor
## 1 Attended grad school
## 2 Attended grad school
## 3       No grad school
```

```
filter(CEO, salary == 358)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
## 1    358  64       1    0     43     11      45    423 Attended college
## 2    358  50       1    1     23      4      25   2300 Attended college
##            grad.factor
## 1       No grad school
## 2 Attended grad school
```

```
filter(CEO, salary == 379)
```

```
##   salary age college grad comten ceoten profits mktval   college.factor
## 1    379  51       1    1      9      3      40   1100 Attended college
## 2    379  55       1    1      4      2      NA     NA Attended college
##            grad.factor
## 1 Attended grad school
## 2 Attended grad school
```

There are 3 salaries with 650,000. They are all different ages. Two of them have the same profits and two of them have the same market value. It's only 3 observations, but this might be something to look into more about where the data came from or if it has been manipulated in some way.

The other salaries (358, 379) don't seem to have any other similarities with each other.

We want to better understand the relationship between salary and profits. We'll plot our points and a least squares regression line to see if there's a relationship.

```
ggplot(na.omit(CEO), aes(x=salary, y=profits)) + geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Profits for different salary levels",
       x = "Salary in thousands", y = "Profits in millions") +
  scale_x_continuous(breaks = seq(0, 5000, 1000)) +
  scale_y_continuous(breaks = seq(-500, 3500, 500))
```

Our regression line shows that there is an overall positive relationship between salary and profits. A higher salary is associated with higher profits. This is not to say that the relationship is linear. It only shows us the best fitting line. There are also several outliers. The highest salary for a CEO is over 5 million, but this company doesn't have anywhere near the highest profits.

The linear regression line doesn't fit very well. That's a lot of noise above it suggesting a different model might be a fit better for this sample.

We can verify the positive relationship by looking at the salary and profit sample correlation.

```
cor(CEO$salary, CEO$profits, use = "complete.obs")
```

```
## [1] 0.3942312
```

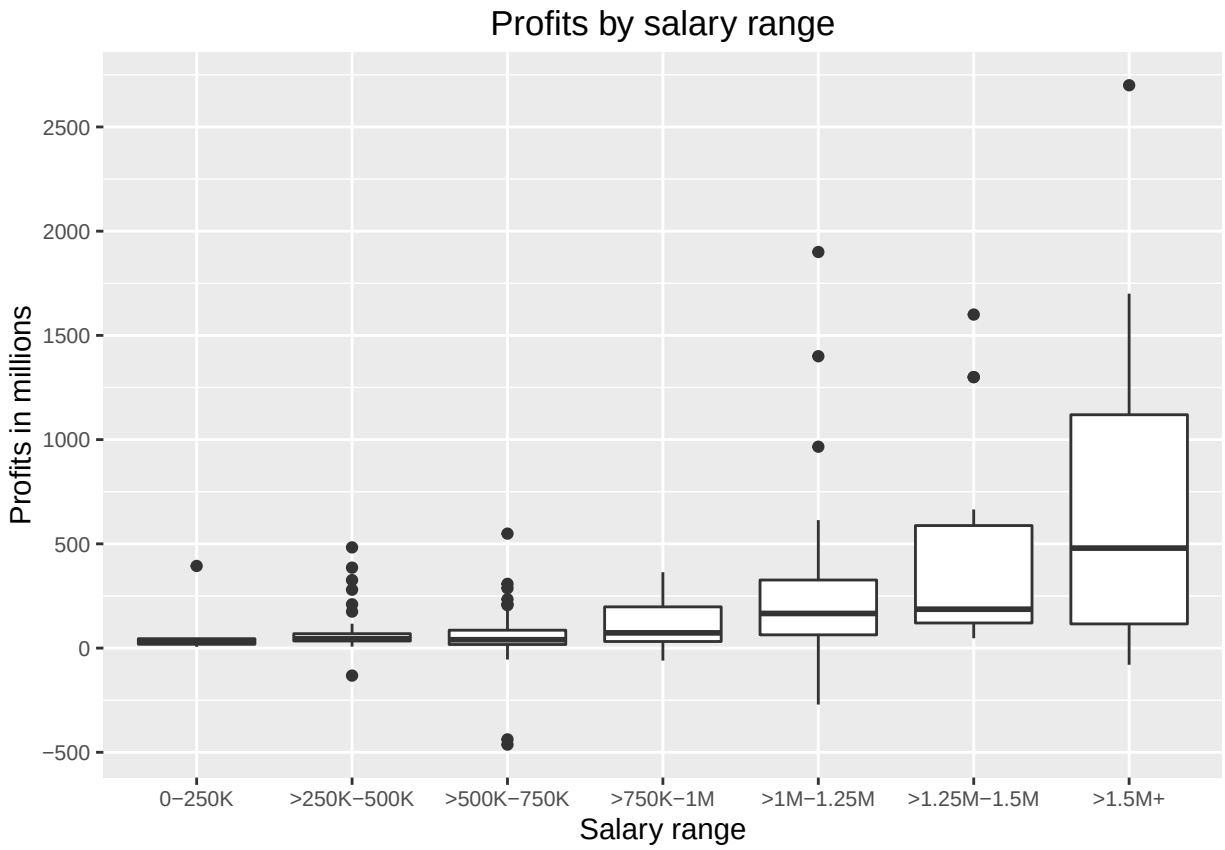
The salary and profits show a positive (.39) correlation, but only moderate in magnitude.

We could also look at the salary and profits boxplots. We will group the salaries into several different buckets.

```
salary_bins = cut(CEO$salary,
                  breaks = c(0, 250, 500, 750, 1000, 1250, 1500, Inf),
                  labels = c("0-250K", ">250K-500K", ">500K-750K",
                             ">750K-1M", ">1M-1.25M", ">1.25M-1.5M",
                             ">1.5M+"))
ggplot(CEO, aes(x = salary_bins, y = profits)) +
  geom_boxplot() +
  labs(title = "Profits by salary range", x = "Salary range",
       y = "Profits in millions") +
```

```
  scale_y_continuous(breaks = seq(-500, 3500, 500)) +
  theme(axis.text=element_text(size=8))
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



From the boxplots we can see that the median profits are generally moving up the more the salary increases. It looks like there is a slight decrease at the 500-750 salary range from the 250-500 salary range, but overall the trend is positive. There's also not a lot of data at the extreme ends of the salary.

We will make plot of salary and market value to see if there's a relationship.

```
ggplot(na.omit(CEO), aes(x=salary, y=mktval)) + geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Market values for different salary levels",
       x = "Salary in thousands", y = "Market value in millions") +
  scale_x_continuous(breaks = seq(0, 5000, 1000)) +
  scale_y_continuous(breaks = seq(0, 50000, 10000))
```

This scatter plot and line shows that the salary has a positive relationship with market value. This is not a linear relationship, but the line indicates a positive correlation. A higher salary is associated with a higher market value. This graph looks very similar to the salary and profits graph. It shows the same highly paid CEO outlier as well.

The linear regression line doesn't look like a good fit here either. There's a lot of noise and high values that don't seem to be captured by the simple model.

Let's look at the correlation between salary and market value.

```
cor(CEO$salary, CEO$mktval, use = 'complete.obs')
```

```
## [1] 0.4066159
```

There seems to be a moderate correlation (.4) between salary and market value.

We will also take a look at salary ranges and market values with some boxplots. We'll use the same salary bins as we used before.

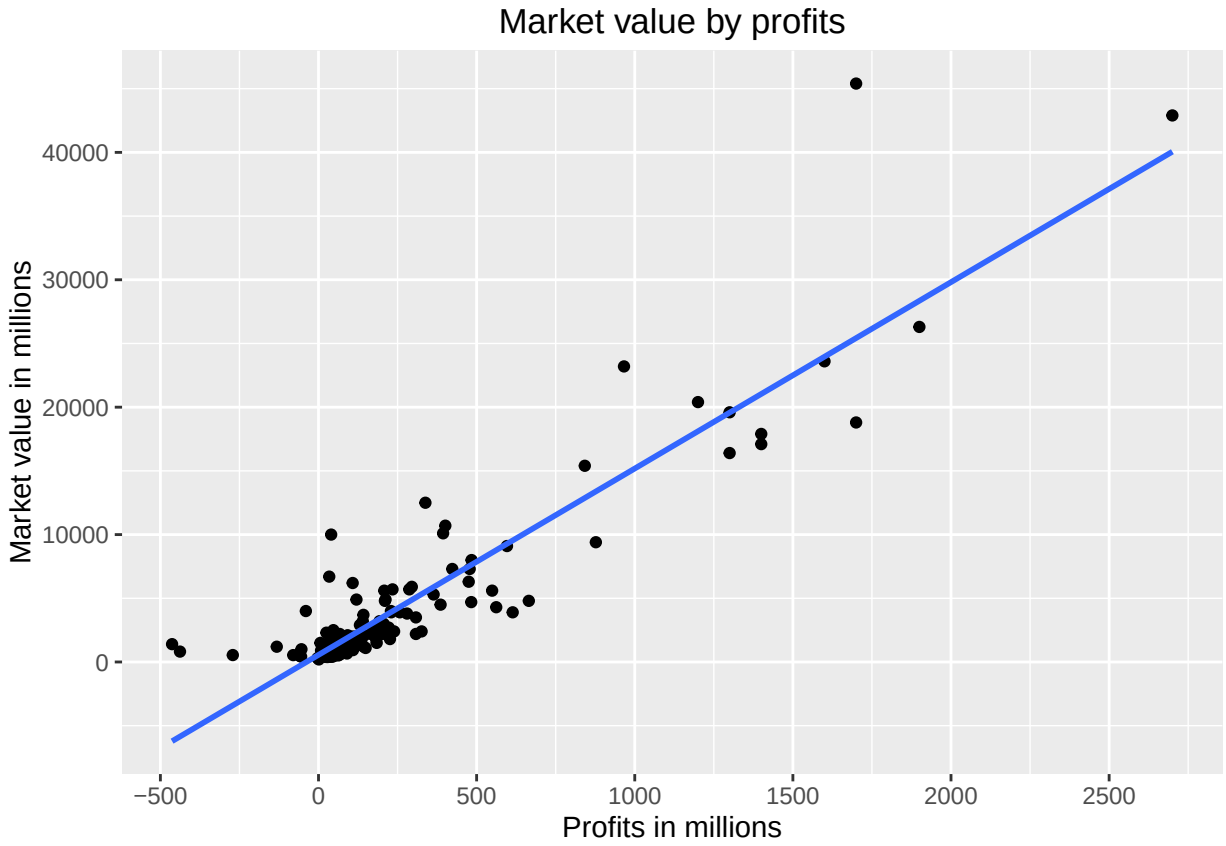
```
ggplot(CEO, aes(x=salary_bins, y=mktval)) +
  geom_boxplot() +
  labs(title = "Market value by salary range", x = "Salary range",
       y = "Market value in millions") +
  scale_y_continuous(breaks = seq(0, 50000, 10000)) +
  theme(axis.text=element_text(size=8))
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

From the boxplots of salary and market value we see a similar trend as the salary and profits boxplots. The median market value is increasing as the salary increases.

We should also explore the relationship between profits and market value.

```
ggplot(na.omit(CEO), aes(x=profits, y=mktval)) + geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Market value by profits", x = "Profits in millions",
       y = "Market value in millions") +
  scale_x_continuous(breaks = seq(-500, 3500, 500)) +
  scale_y_continuous(breaks = seq(0, 50000, 10000))
```



The graph of market value and by profits shows a strong relationship between profits and market value. This is expected since profit is used to measure the value of the company. We can also see that companies with negative profit still had some market value to them which the regression line doesn't capture. If we were only concerned about market value, then profits might be a better indicator than salary. Since were interested in salary and company performance there could be some confounding with salary, profits and market value.

We can look at the correlation too to see how strong the relationship is between profits and market value.

```
cor(CEO$profits, CEO$mktval, use = "complete.obs")
```

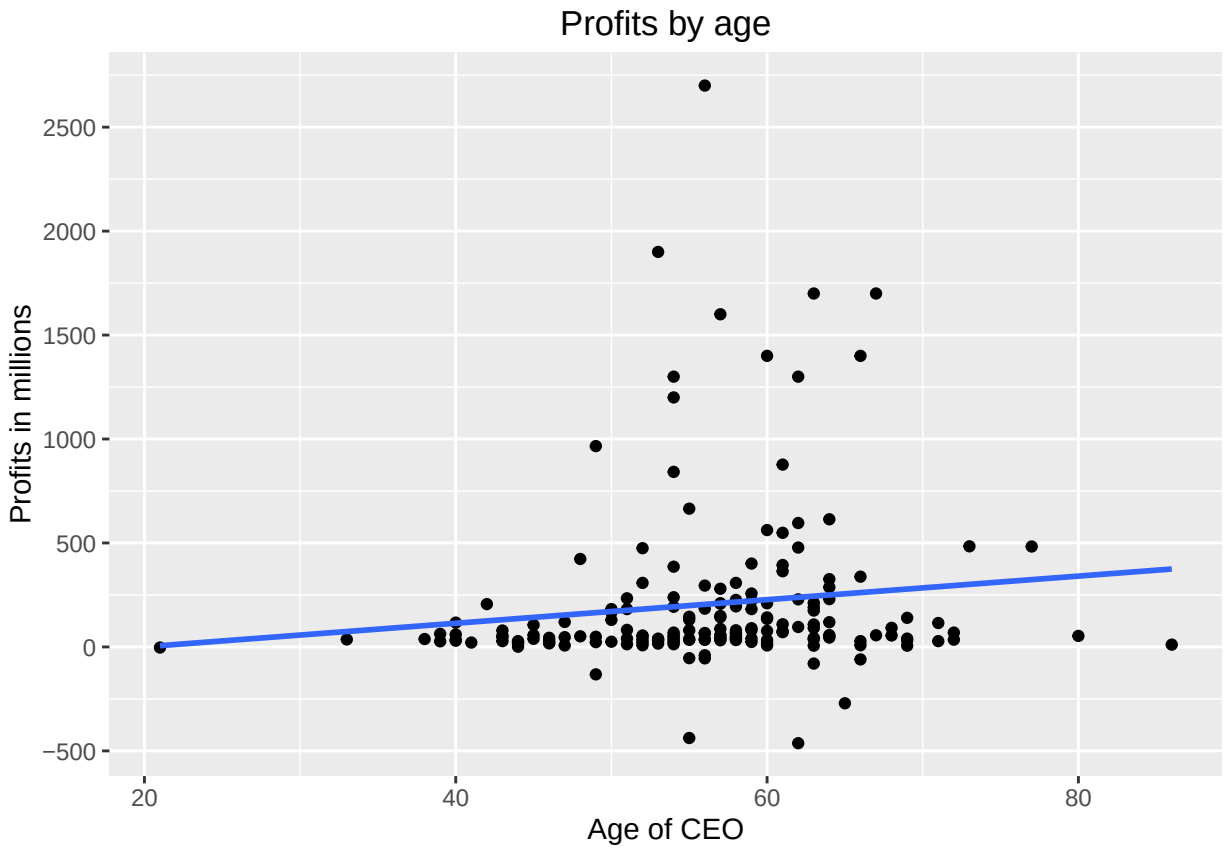
```
## [1] 0.9183732
```

The correlation (.91) shows strong correlation between profits and market value. This is much higher than the correlation between the salary and the profits we looked at earlier and the correlation between salary and market value.

While salary is our main focus for this exploration we also want to look at age, time at company, college, grad school and how those relate to profit and market value to see if we find any other valuable relationships.

We'll first look at age and profits in a scatter plot.

```
ggplot(na.omit(CEO), aes(x = age, y = profits)) + geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Profits by age", x = "Age of CEO",
       y = "Profits in millions") +
  scale_y_continuous(breaks = seq(-500, 3500, 500))
```



From this graph we can see a very slight positive relationship from the best fit line. This indicates an association between older CEOs and higher profits. However, the linear regression line is not a very good fit. There's a lot of noise above the line and perhaps an inverted parabola would be a better fit for this sample.

Let's look at the correlation between age and profit.

```
cor(CEO$age, CEO$profits, use = "complete.obs")
```

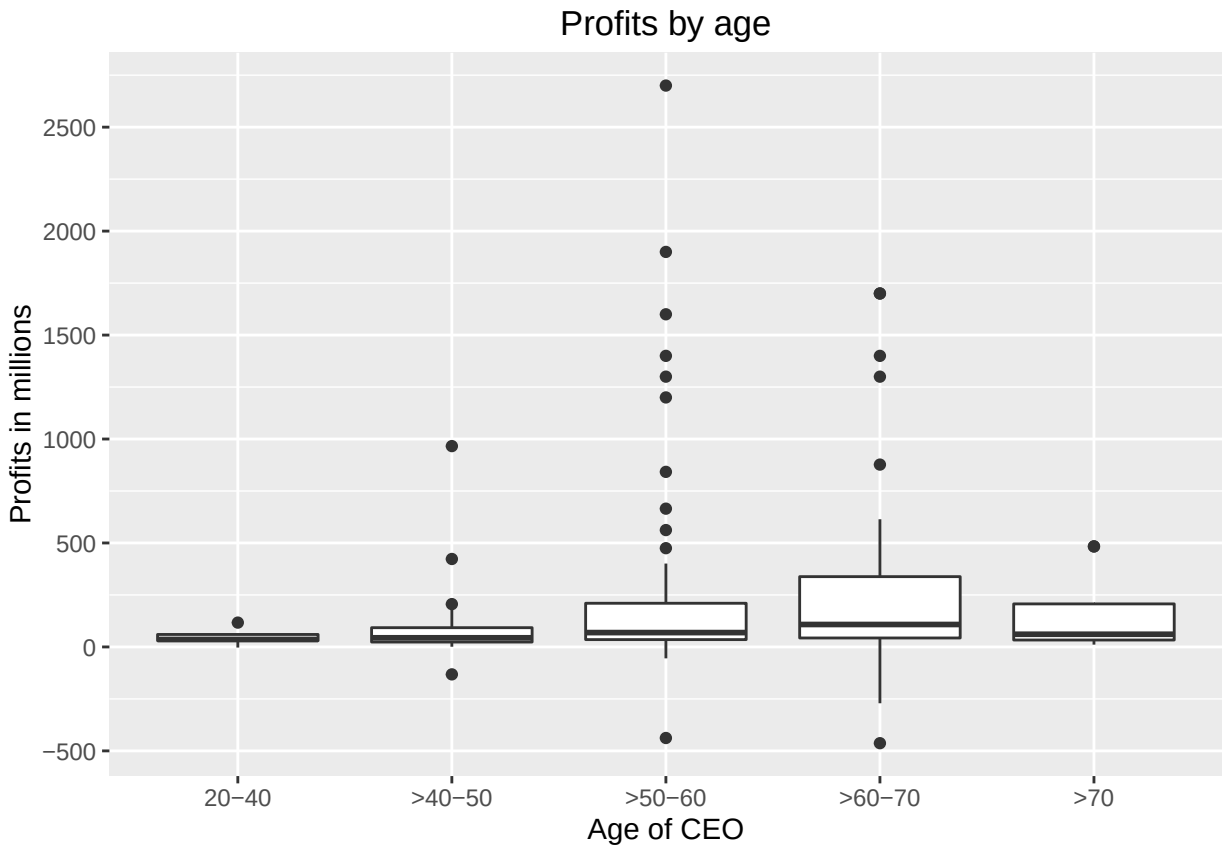
```
## [1] 0.124336
```

This correlation between age and profit is pretty low (.12).

Let's make some boxplots of age and profits.

```
age_bins = cut(CEO$age,
                  breaks = c(20, 40, 50, 60, 70, Inf),
                  labels = c("20-40", ">40-50", ">50-60",
                             ">60-70", ">70"))
ggplot(CEO, aes(x = age_bins, y = profits)) + geom_boxplot() +
  labs(title = "Profits by age", x = "Age of CEO",
       y = "Profits in millions") +
  scale_y_continuous(breaks = seq(-500, 3500, 500))
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

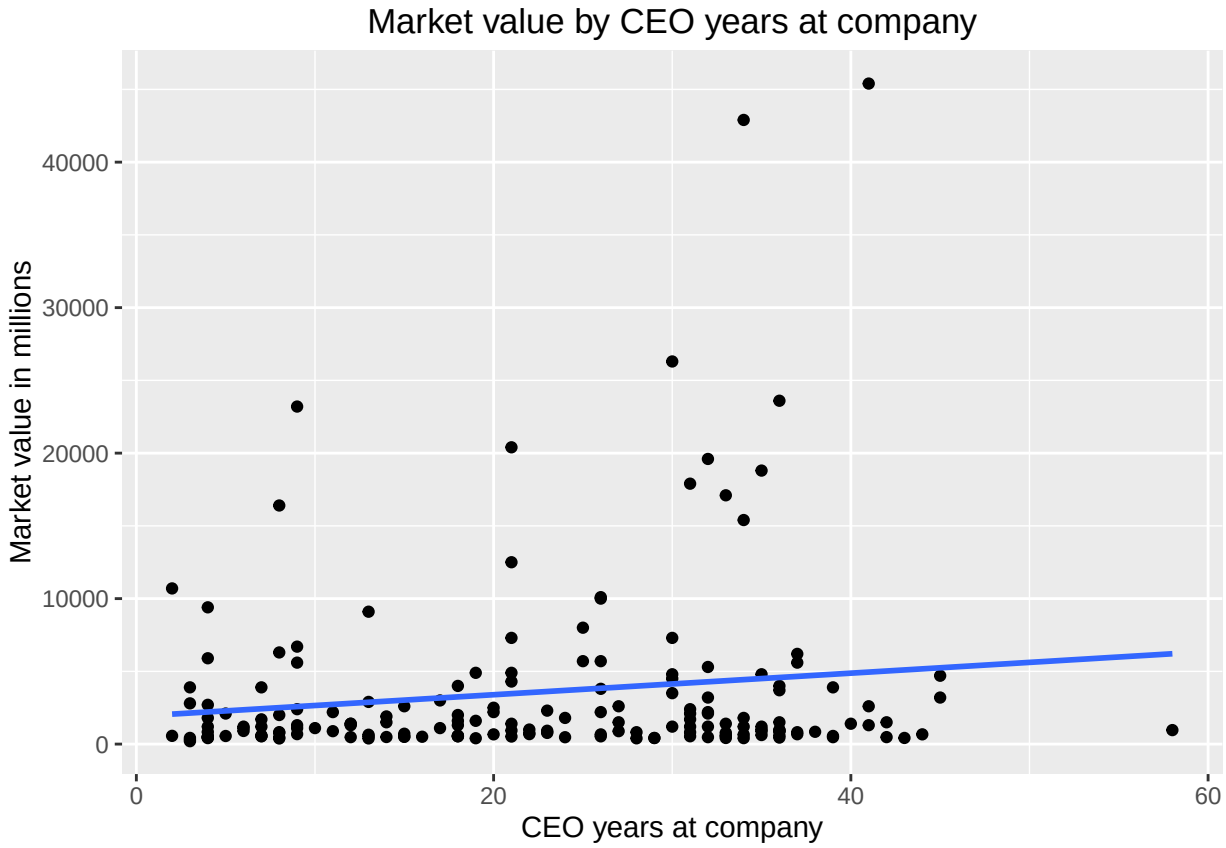


These boxplots show the same small positive relationship between the age groups and median profits. The age group for >70 starts to go back down. This is likely due to few people at that age still working. The association between age and profits could mean that older people perform better at the CEO position, but the relationship is very weak and is not a good indicator.

Overall, the scatter plot and boxplots of age don't appear to have a strong relationship with profits. We will not explore age and market value since the scatterplot indicated that there was not a strong relationship earlier.

We will now take a look at years at the company and market value.

```
ggplot(na.omit(CEO), aes(x = comten, y = mktval)) + geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Market value by CEO years at company",
       x = "CEO years at company",
       y = "Market value in millions") +
  scale_y_continuous(breaks = seq(0, 50000, 10000))
```



We see that the amount of time at a particular company has a very small positive relationship with the market value. This could also be that companies grow over time and not, because the person was at the company for many years. The linear regression line here doesn't fit very well either. It has a lot of noise above it.

Let's calculate the correlation between years at company and market value.

```
cor(CEO$comten, CEO$mktval, use = "complete.obs")
```

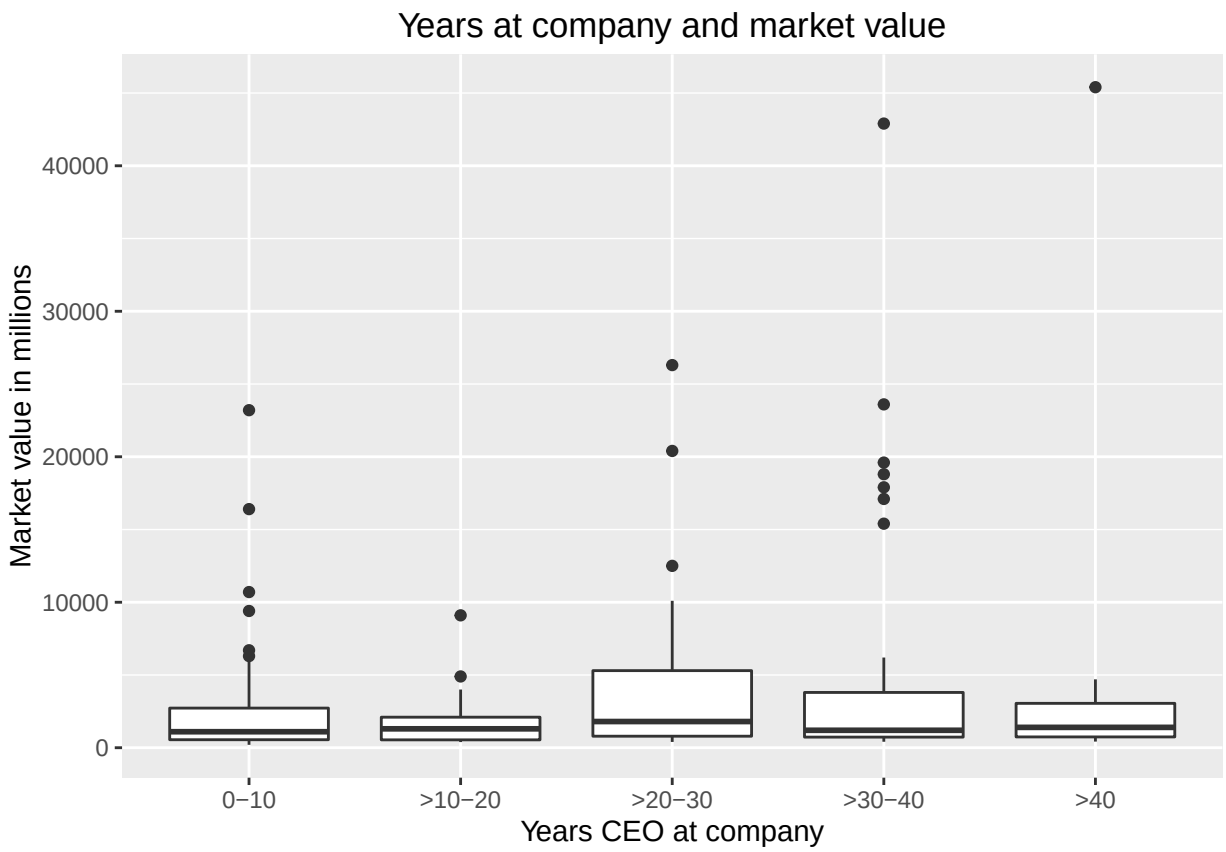
```
## [1] 0.1431319
```

We see a small correlation (.14) between years at company and market value. This is likely not a very good indicator of company performance.

We'll create some boxplots for years at company and market value.

```
years_at_bins = cut(CEO$comten,
                    breaks = c(0, 10, 20, 30, 40, Inf),
                    labels = c("0-10", ">10-20", ">20-30",
                               ">30-40", ">40"))
ggplot(CEO, aes(x = years_at_bins, y = mktval)) +
  geom_boxplot() +
  labs(title = "Years at company and market value",
       x = "Years CEO at company",
       y = "Market value in millions") +
  scale_y_continuous(breaks = seq(0, 50000, 10000))
```

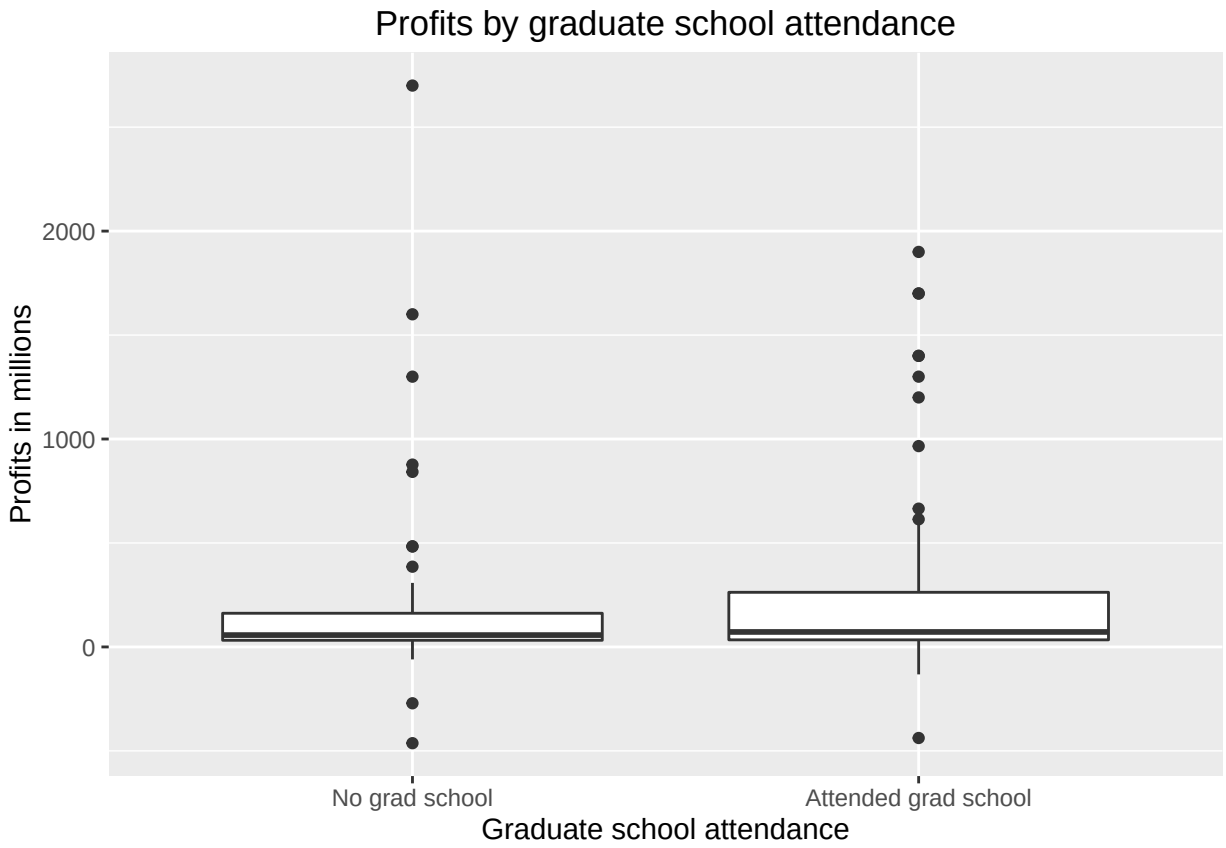
```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



When the ages are broken into groups we see a different trend. The median for years at company goes up slightly towards the middle and then down and then slightly up again. These are all small differences though and are not indicators of any strong relationships.

We should also look at graduate school attendance to see if higher education has any association with company performance.

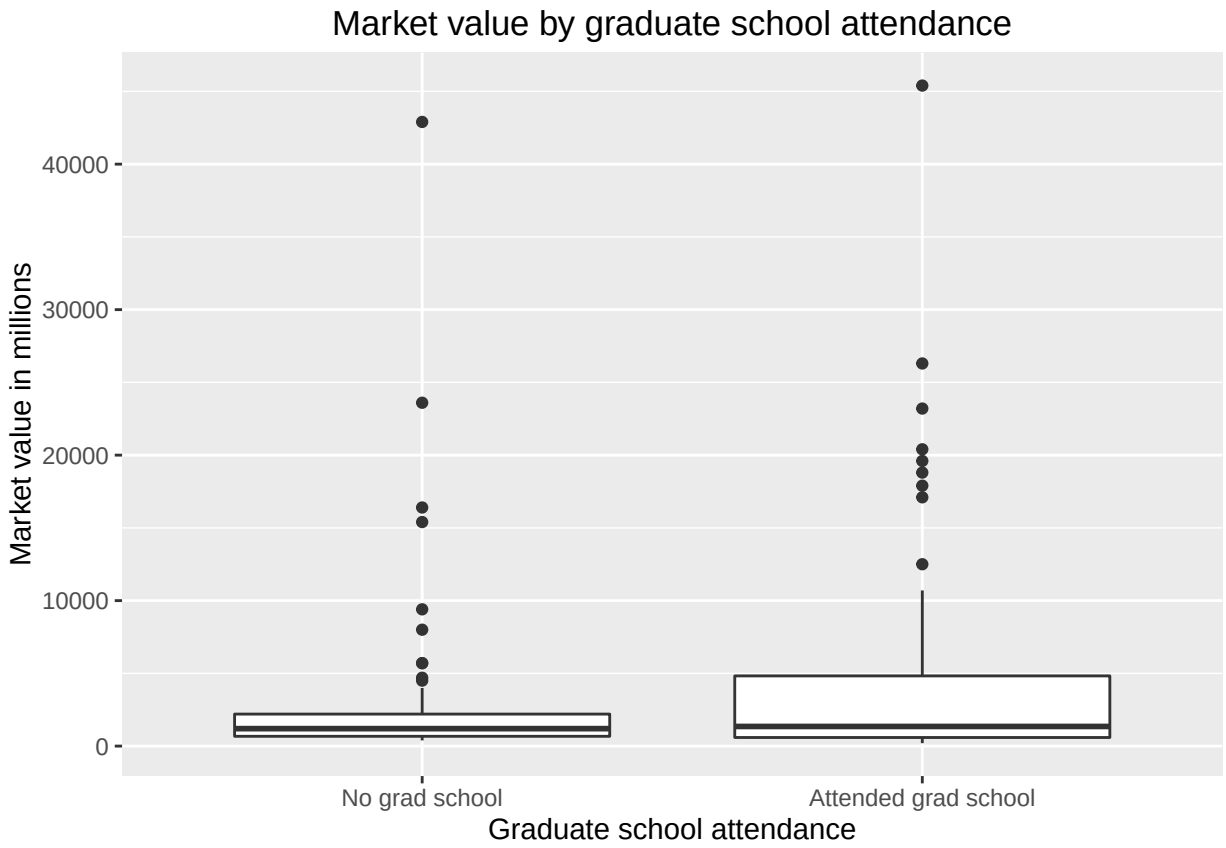
```
ggplot(na.omit(CEO), aes(x = grad.factor, y = profits)) +
  geom_boxplot() +
  labs(title = "Profits by graduate school attendance",
       x = "Graduate school attendance",
       y = "Profits in millions")
```



The median profits are slightly higher for attending grad school. This doesn't show a big difference between profits and grad school attendance. It's likely this variable is not a good predictor of profits.

We'll also look at market value and graduate attendance.

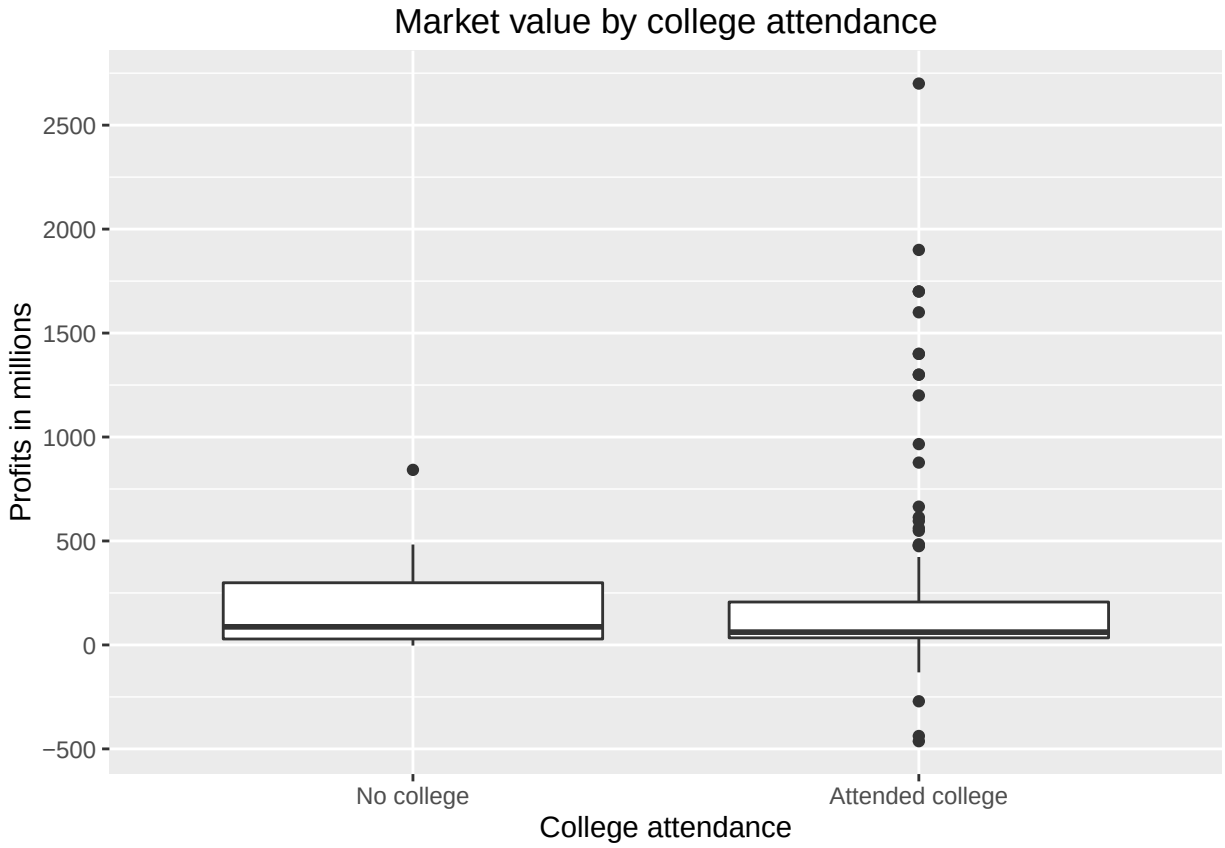
```
ggplot(na.omit(CEO), aes(x = grad.factor, y = mktval)) + geom_boxplot() +
  labs(title = "Market value by graduate school attendance",
       x = "Graduate school attendance",
       y = "Market value in millions")
```



This shows a similar relationship as graduate school attendance and profits. Here the median market value for CEOs that attended graduate school is only slightly higher that for those without graduate school. This seems to indicate that graduate school is not a good indicator of company performance.

We should also check college attendance.

```
ggplot(na.omit(CEO), aes(x = college.factor, y = profits)) +
  geom_boxplot() +
  labs(title = "Market value by college attendance",
       x = "College attendance",
       y = "Profits in millions") +
  scale_y_continuous(breaks = seq(-500, 3500, 500))
```



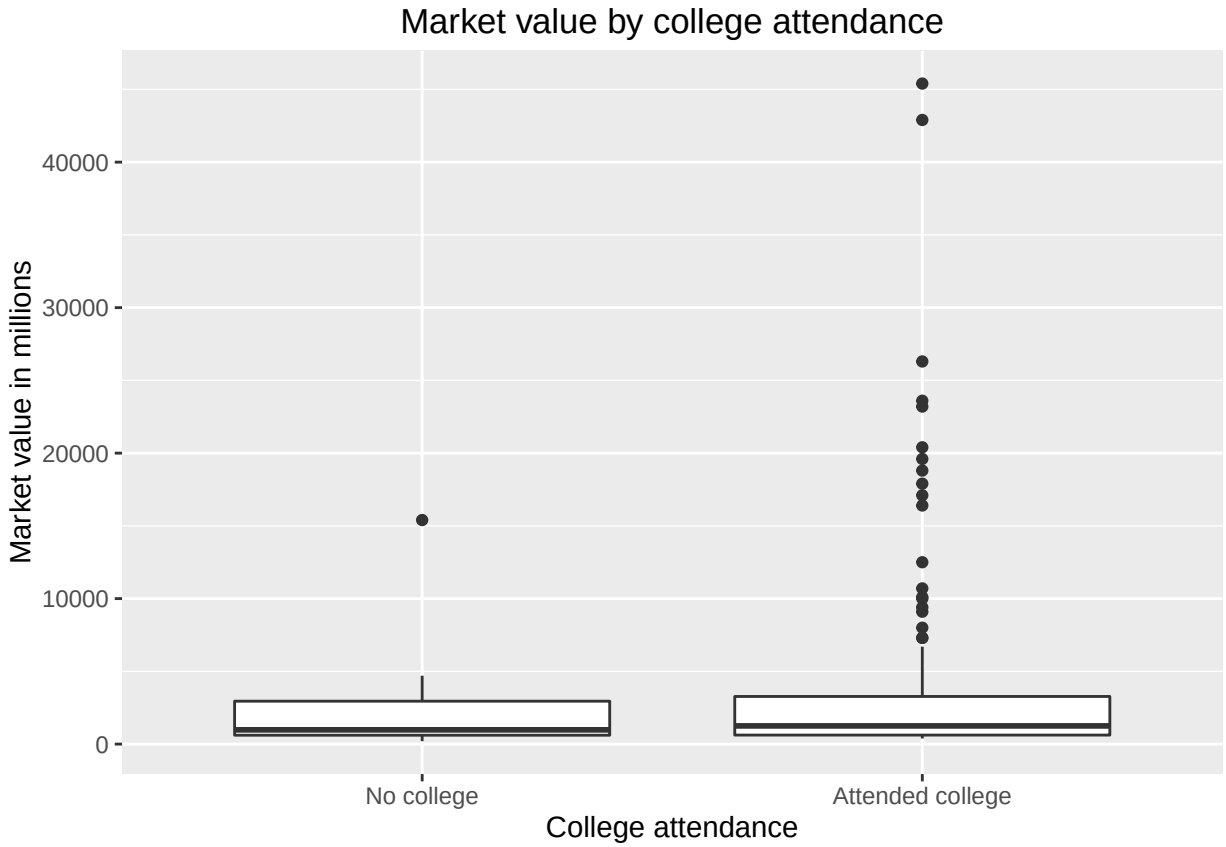

```
(nrow(CEO[CEO$college == 0,]))
```

```
## [1] 7
```

Here we see that not attending college has a slightly higher median value. However there were only a few observations in the dataset that had no college (7). It would be hard to draw conclusions from such a small number.

Let's see if the same is true for market value and college.

```
ggplot(na.omit(CEO), aes(x = college.factor, y = mktval)) + geom_boxplot() +
  labs(title = "Market value by college attendance",
       x = "College attendance",
       y = "Market value in millions") +
  scale_y_continuous(breaks = seq(0, 50000, 10000))
```



This shows a different relationship. Those that attended college have a median market value slightly higher than those that did not attend college. While interesting there's only 7 observations for those not attending college and it doesn't show a strong relationship between attending college and market value.

Discussion

We looked how salary is associated with profits and market value. We've shown that there is a moderate association between salary and profits and salary and market value. Salary would be a good feature to use for statistical modeling. We also saw that profit has a very strong association with market value and while it was expected it may be useful to add it to our statistical models. There were a lot of outliers and positive skew in our data. The linear regression lines didn't always capture our data very well and perhaps other models should be considered.

While our focus was on salary we saw that age, years at company, college, and grad school had small associations with company performance and could affect our statistical analysis.

Early in the discovery process we uncovered several observations that couldn't be used due to the same negative values in market value and profits. We then omitted those 5 observations from our exploration analysis. The small number likely didn't affect the relationships that we discovered. It might be worth checking with the source of the data about these values. We also saw several duplicate values in salary, market value, and profits. While we didn't throw them out it was worth noting that these values might have an effect on future analysis.