W203 Statistics - Lab 1

Justine Heritage, Morris Burkhardt, Daniel Volk

May 31, 2017

Introduction

This analysis is motivated by the following research question:

> What is the relationship between CEO salary and company performance?

Our data was provided with the following codebook:

Variable Name	Variable Meaning
salary	1990 compensation, $1000s
age	in years
college	=1 if attended college
grad	=1 if attended graduate school
comten	years with company
ceoten	years as CEO with company
profits	1990 profits, millions
mktval	market value, end 1990, millions

As we look at these variables, we need to define how we will measure company performance and consider possible limitations in this dataset.

Our primary indicators of company performance will be profit and market value. However, there are several other factors than CEO salary that could influence a company's performance. For example:

- Performance of the CEO's predecessor
- Market environment of a company's industry or sector
- Performance of executive team and employees
- Influence from board of directors
- Prior year profits

To address some of these issues, we will also examine the length of the CEO's tenure. However, a lack of knowledge about external factors will prevent us from making any claims of causality.

Setup

To begin our analysis, we used the `car` library and loaded the provided data set.

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

We take an initial look at our data set.

```
head(CEO)
```

```
##     salary age college grad comten ceoten profits mktval
## 154   1033  62       1    1     30      1     478   7300
## 79     879  63       1    1     21      9     212   4900
```

```
## 19     971  72       1    1     33     24      69    609
## 115    567  56       1    0     31     10      65   1700
## 36    1336  60       1    1     21     13     562   4300
## 153   1444  59       1    1      2      2     401  10700
```

```
names(CEO)
```

```
## [1] "salary"  "age"     "college" "grad"    "comten"  "ceoten"  "profits"
## [8] "mktval"
```

```
str(CEO)
```

```
## 'data.frame':    185 obs. of  8 variables:
##  $ salary : num  1033 879 971 567 1336 ...
##  $ age    : num  62 63 72 56 60 59 46 59 51 56 ...
##  $ college: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ grad   : num  1 1 1 0 1 1 1 1 0 1 ...
##  $ comten : num  30 21 33 31 21 2 7 3 8 9 ...
##  $ ceoten : num  1 9 24 10 13 2 3 3 8 3 ...
##  $ profits: num  478 212 69 65 562 401 44 257 13 34 ...
##  $ mktval : num  7300 4900 609 1700 4300 10700 533 3900 458 6700 ...
```

There are 185 observations over 8 variables. We notice that `college` and `grad` are dummy variables. The rest of the variables are numeric. The `salary` variable is measured in millions of \$, `profits` and `mktval` are measured in thousands of \$. The `age`, `comten` (years the CEO has been with the company) and `ceoten` (years as CEO with the company) variables are integer values.

Data Selection

We summarize the data.

```
summary(CEO)
```

```
##      salary            age           college            grad
##  Min.   : 100.0   Min.   :21.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 467.0   1st Qu.:51.00   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 697.0   Median :57.00   Median :1.0000   Median :1.0000
##  Mean   : 852.9   Mean   :55.78   Mean   :0.9622   Mean   :0.5514
##  3rd Qu.:1101.0   3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5299.0   Max.   :86.00   Max.   :1.0000   Max.   :1.0000
##      comten          ceoten          profits           mktval
##  Min.   : 2.00   Min.   : 0.000   Min.   :-463.0   Min.   :   -1
##  1st Qu.: 9.00   1st Qu.: 3.000   1st Qu.:  33.0   1st Qu.:  567
##  Median :21.00   Median : 5.000   Median :  57.0   Median : 1200
##  Mean   :21.66   Mean   : 7.681   Mean   : 199.2   Mean   : 3450
##  3rd Qu.:33.00   3rd Qu.:11.000   3rd Qu.: 195.0   3rd Qu.: 3200
##  Max.   :58.00   Max.   :37.000   Max.   :2700.0   Max.   :45400
```

We notice that there is an unusual minimum, -1, for `mktval`. We determine that in both `mktval` and `profits`, these observations have missing data. We replace the -1s with `NA`.

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

By looking at the tail of our data, we also notice that row 184 and 185 might be duplicates. All values - besides age - are identical. These two rows are also the only rows that indicate that the CEO went to grad school but not college. It's possible that this data is inconsistent.

```r
# final 2 values look duplicated

tail(CEO[order(as.numeric(row.names(CEO))),],2)
```

```
##     salary age college grad comten ceoten profits mktval
## 184    453  33       0    1      3      1      33    344
## 185    453  30       0    1      3      1      33    344
```

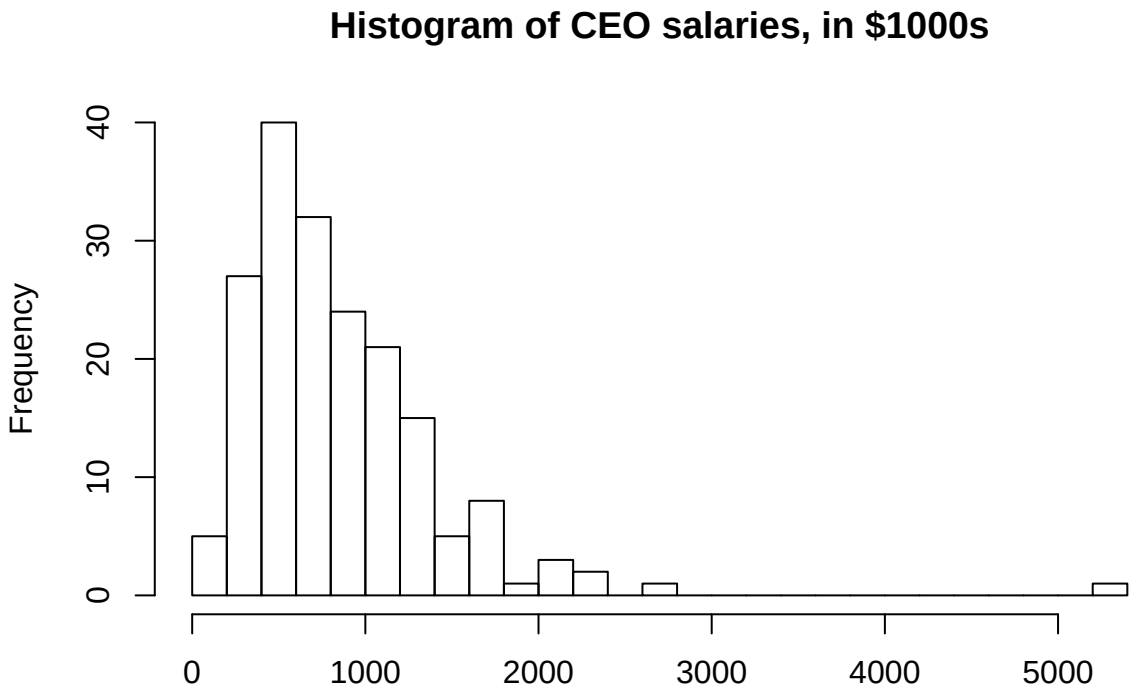
```r
# potentially synthesize rows into 1 with age value 31.5
```

Exploratory Analysis

Univariate Analysis

First, we look at the histogram for `salary`.

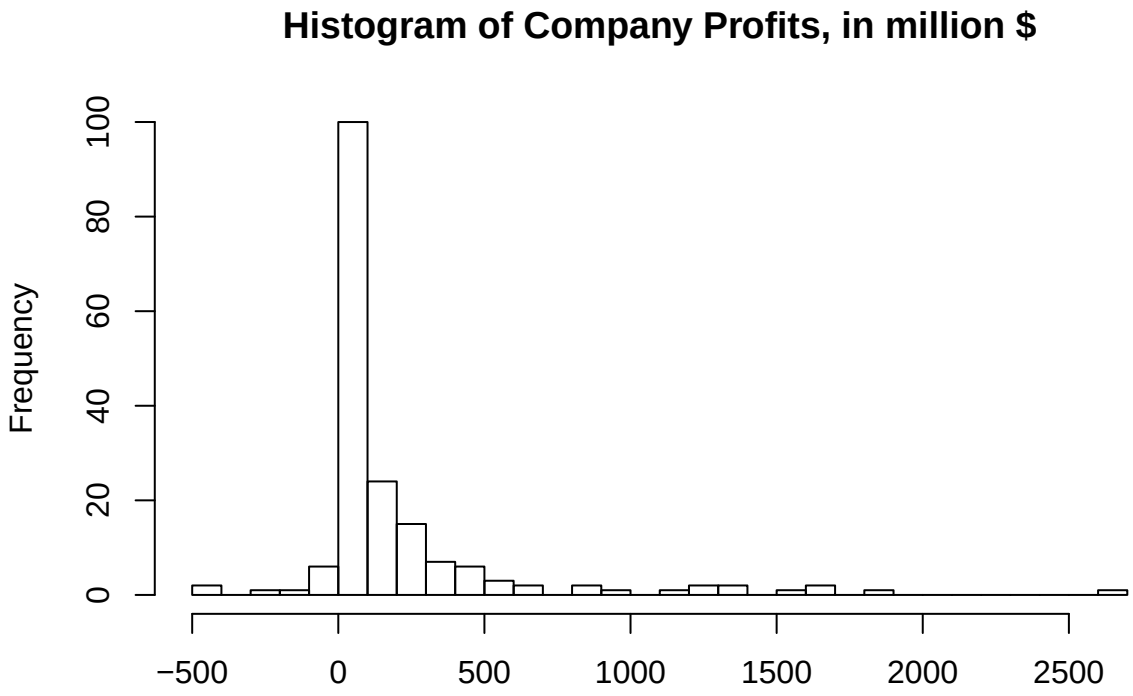
```r
hist(CEO$salary, breaks = 20, main = "Histogram of CEO salaries, in $1000s" ,
     xlab = NULL)
```



We notice that the data is positively skewed and there are some large outliers.

Next, we look at the histogram for `profits`.

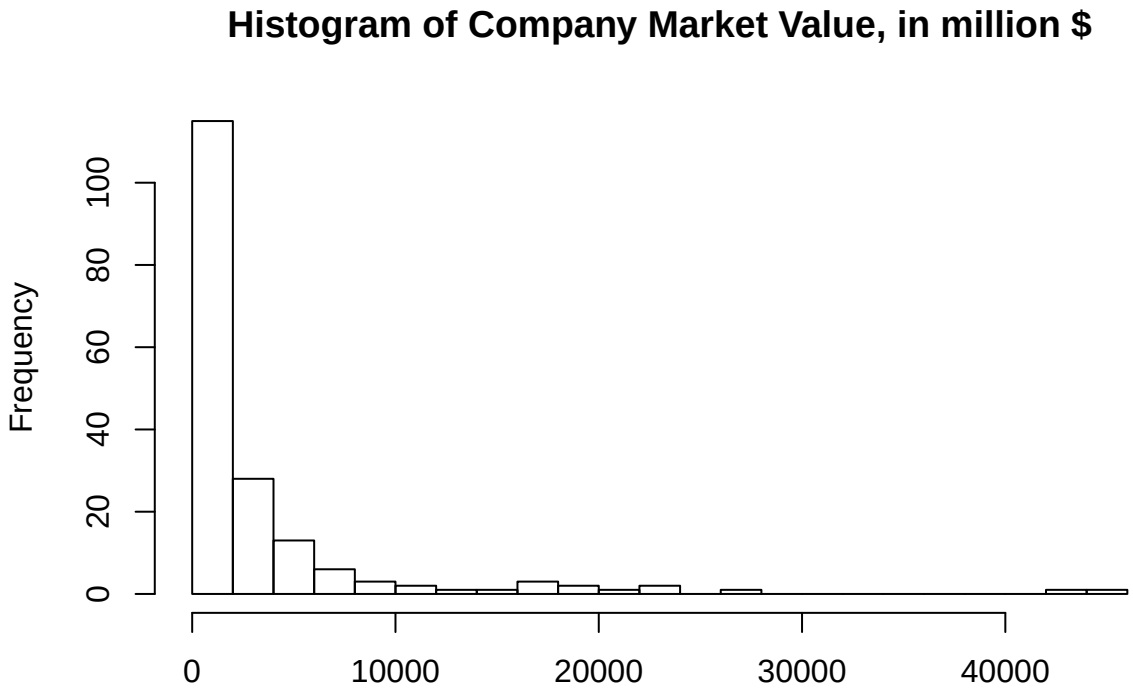
```r
hist(CEO$profits, breaks = 30, main = "Histogram of Company Profits, in million $" ,
     xlab = NULL)
```



The `profits` variable has both negative values and outliers at the high end.

Finally, we look at the histogram form `mktval`.

```
hist(CEO$mktval, breaks = 20, main = "Histogram of Company Market Value, in million $" ,
     xlab = NULL)
```



Again, we see large outliers in the `mktval` variable.

Since all distributions are heavily skewed with outliers to the far right, we perform a logarithmic transformation. For the `profits` variable, we also have to consider omitting all values $<= 0$ in the transformation and then analysing those values seperately.

First we check how many profits values actually are zero or negative.

```
sum(CEO$profits<=0, na.rm = TRUE)
```

```
## [1] 10
```

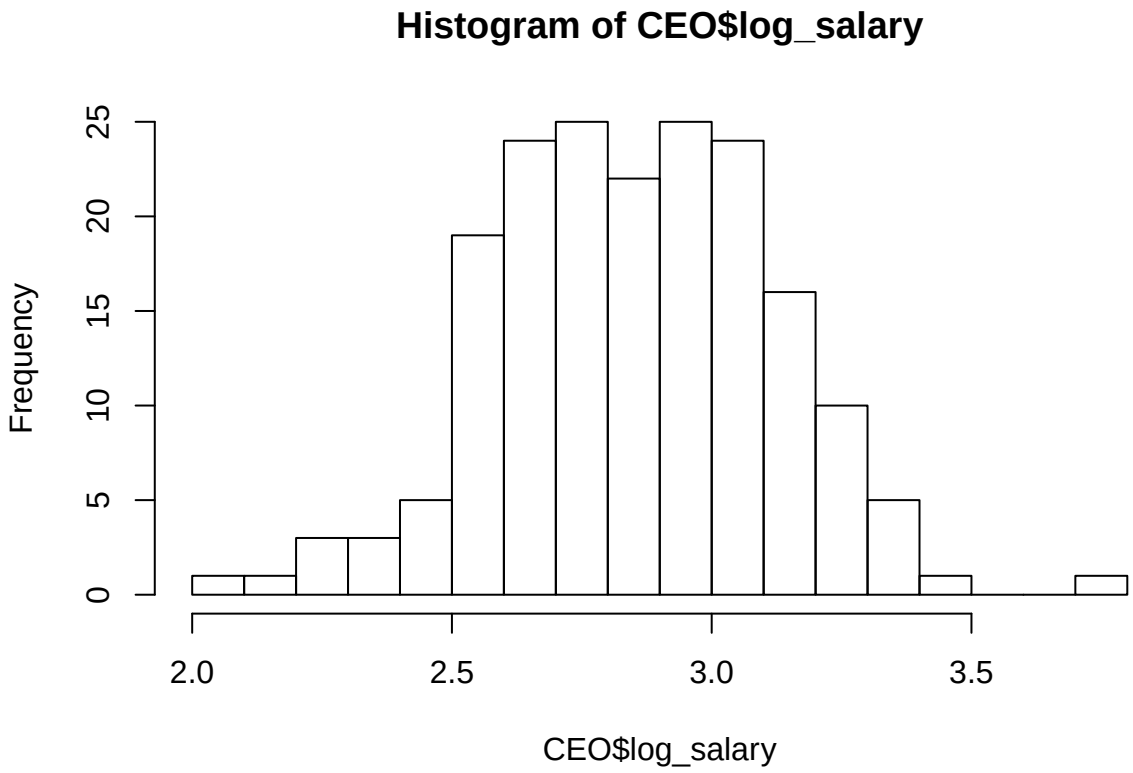
We would be omitting 10 values out of 185, which is less than 6% of our data.

Next, we look at histograms for our log-transformed data.

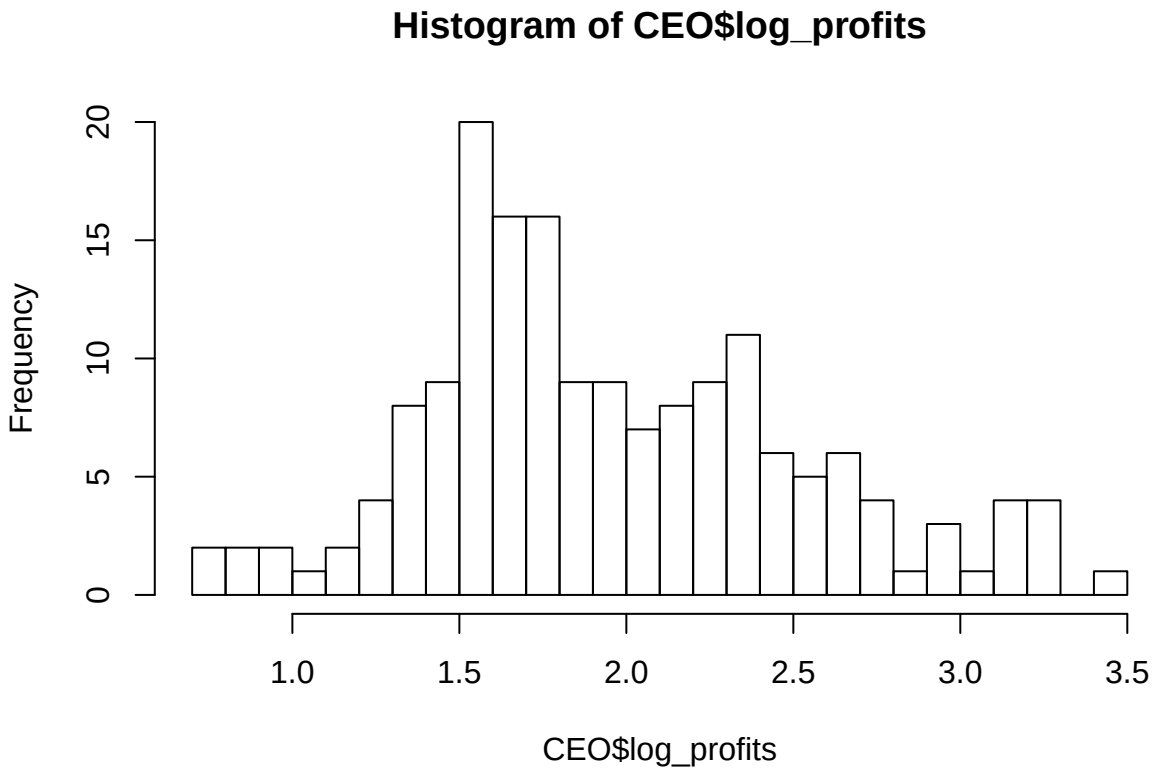
```
CEO$log_salary  = log10(CEO$salary)
CEO$log_profits = log10(CEO$profits)
```

```
## Warning: NaNs produced
```

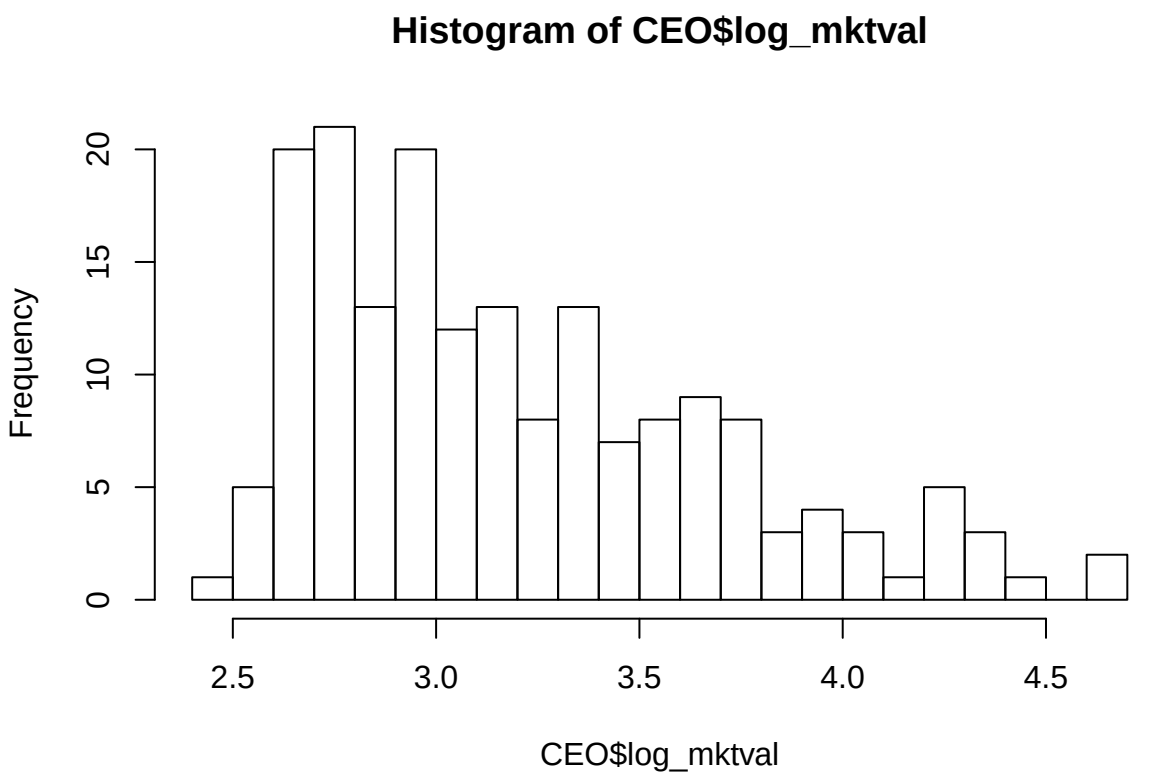
```
CEO$log_mktval  = log10(CEO$mktval)

hist(CEO$log_salary, breaks = 20)
```




```
hist(CEO$log_profits, breaks = 20)
```




```
hist(CEO$log_mktval, breaks = 20)
```



The negative profit values are transformed into NAs.

The transformed `salary` and `profits` variables resemble the normal distribution. The `mktval` variable still has a minor left skew, but with no significant outliers.

Bivariate analysis

Salary, Market Value and Profits

First, we look at the linear correlations between `salary` and `profits` and `salary` and `mktval`. To better compare the two correlation coefficients, we'll only look at positive profit values. Then, we look at a scatterplot matrix for these three variables.

```
pos_profits = CEO$profits > 0

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

```
## [1] 0.4126423
```

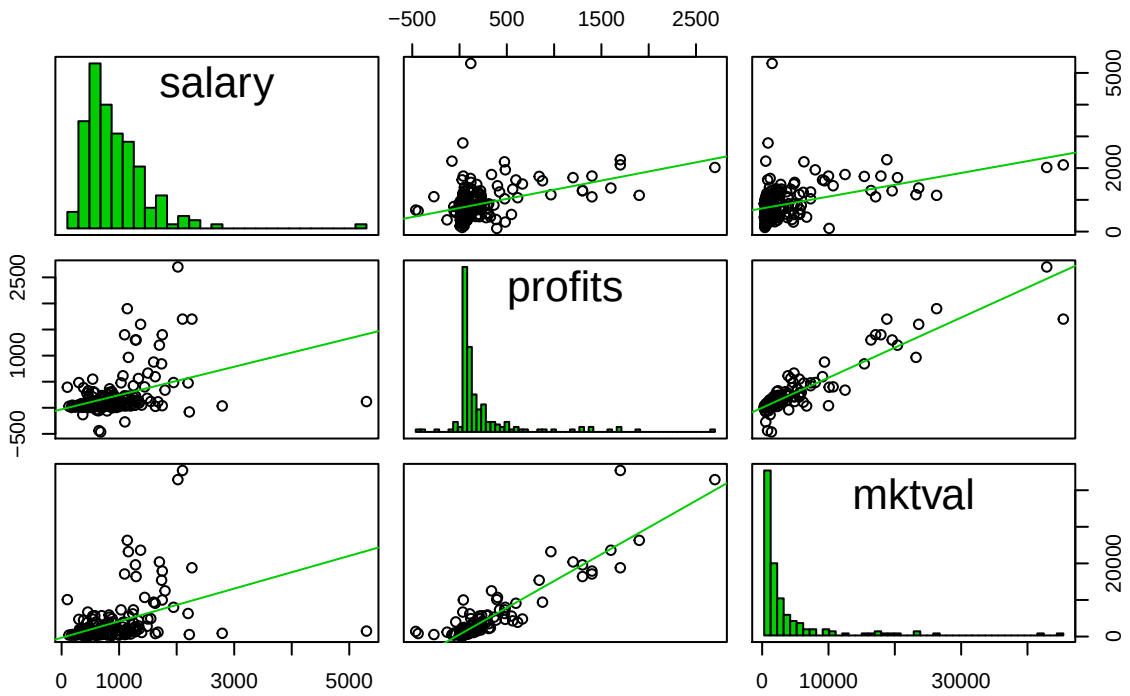
```
cor(CEO$mktval, CEO$salary, use = "complete.obs")
```

```
## [1] 0.4082068
```

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

```
## [1] 0.9293305
```

```
scatterplotMatrix( ~ salary + profits + mktval, data = CEO,
 #                 reg.line="" ,
                   smoother="",
                   diagonal = "histogram",
                   main = "Scatterplot Matrix for original data")
```



There is a moderate positive linear correlation between both `salary` and both `profits` and `mktval`. We also notice a strong linear correlatio between `profits` and `mktval`.

Next we'll look at the correlations between the log-transformed variables and their scatterplot matrix.

```
cor(CEO$log_profits, CEO$log_salary, use = "complete.obs")
```

```
## [1] 0.4775787
```

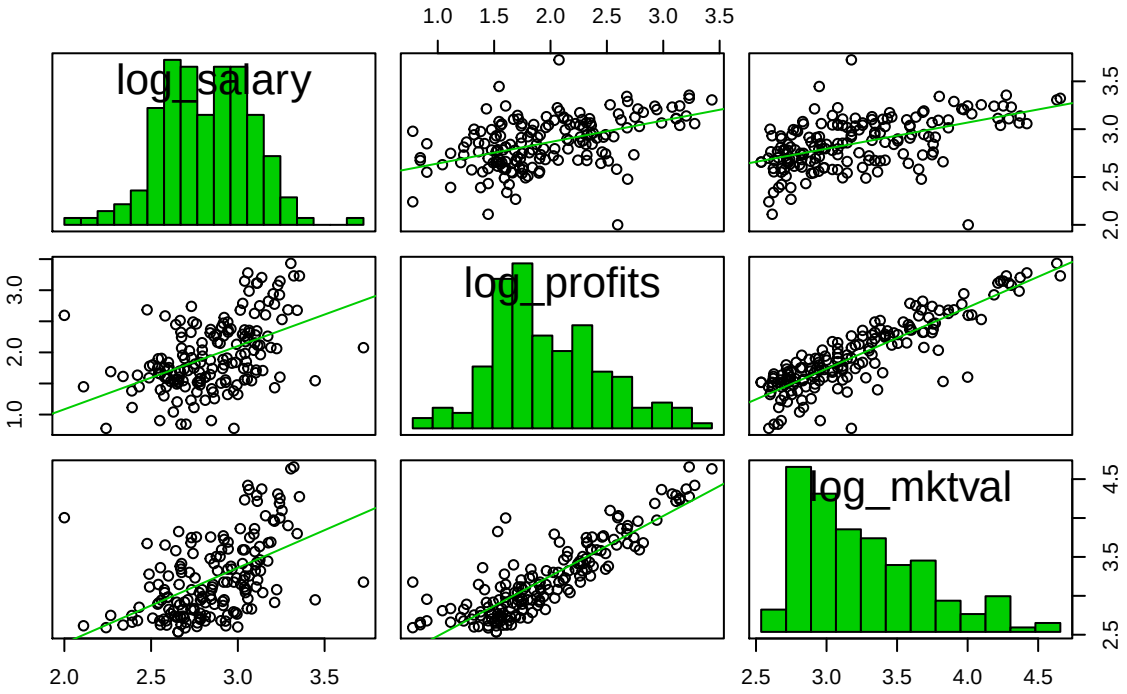
```
cor(CEO$log_mktval, CEO$log_salary, use = "complete.obs")
```

```
## [1] 0.4809051
```

```
cor(CEO$log_profits, CEO$log_mktval, use = "complete.obs")
```

```
## [1] 0.8719542
```

```
scatterplotMatrix( ~ log_salary + log_profits + log_mktval, data = CEO,
 #                 reg.line="" ,
                   smoother="",
                   diagonal = "histogram",
                   main = "Scatterplot Matrix for transformed data")
```



We can see a stronger relationship between `salary` and `profits` as well as `salary` and `mktval` with the correlation increasing from 0.4126423 to 0.4775787 and 0.4082068 to 0.8719542, respectively. Due to the nonlinearity of the relationship, the calculated correlation on the original (non transformed) variables, underestimates the actual relationship.

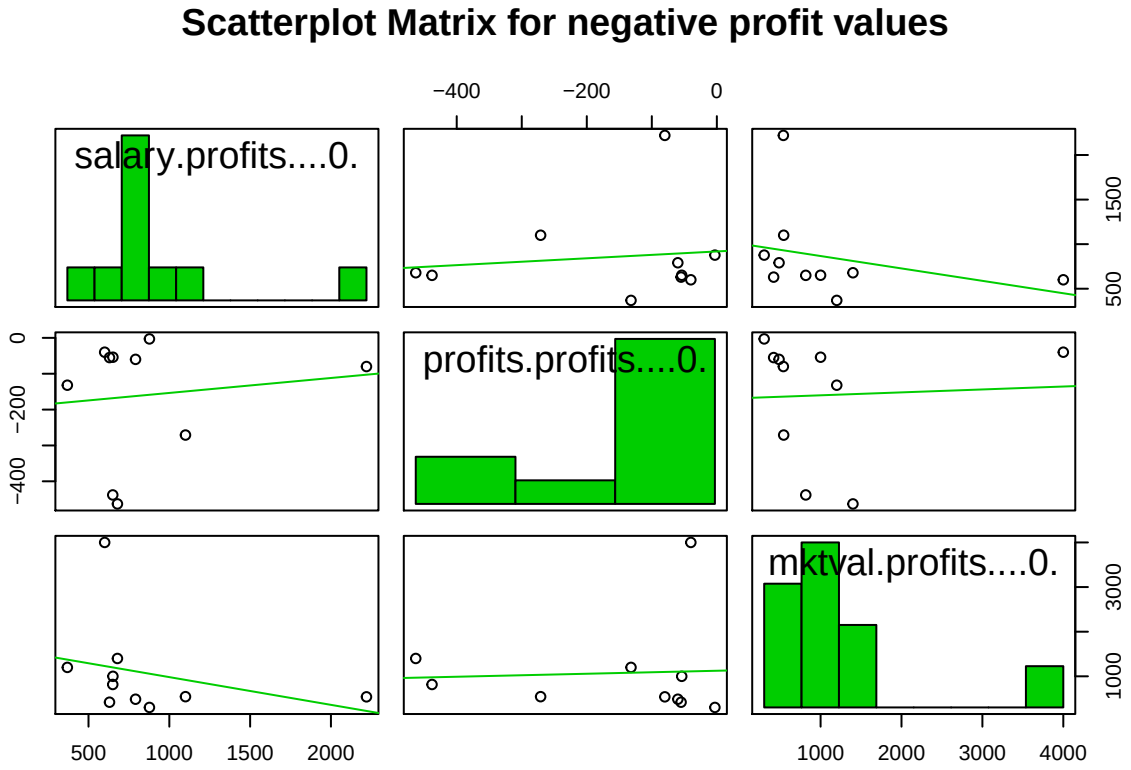
Let's now take a look at the Scatterplot Matrix for the key variables under examination. Note that all data points, where profits is negative, are omitted.

We can see a correlation between

Does it make sense to even look at those ten data points?

```
scatterplotMatrix( ~ salary[profits <= 0] + profits[profits <= 0] + mktval[profits <= 0],
                   data = CEO,
 #                 reg.line="" ,
```

```
                smoother="",
                diagonal = "histogram",
                main = "Scatterplot Matrix for negative profit values")
```



Further considerations

Weighted profits / market value

Maybe it would be better to weight the profits and market value variables with the time the CEO has been CEO respectively with the company. This way we limit the effect of predecessors policies.

```
CEO$weighted_profits = CEO$profits * CEO$ceoten
CEO$weighted_mktval = CEO$mktval * CEO$ceoten

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

```
## [1] 0.3305032
```

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

```
## [1] 0.3127059
```

```
cor(CEO$weighted_profits, CEO$weighted_mktval, use = "complete.obs")
```

```
## [1] 0.9163706
```

Since there are comten/ceoten values that are zero, some of the weighted values are now zero. We replace those with NAs before logarithmic transformation, to avoid infintive values. These values will be omitted and this might be justified by arguing, that CEOs who have been with the company for less than a year, do not have a significant effect on the companies profit or market value yet.

```r
CEO$log_weighted_profits = log10(CEO$weighted_profits)
```

```
## Warning: NaNs produced
```

```r
CEO$log_weighted_mktval  = log10(CEO$weighted_mktval)

CEO$log_weighted_profits[CEO$weighted_profits == 0] = NA
CEO$log_weighted_mktval[CEO$weighted_mktval == 0] = NA

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

```
## [1] 0.4726666
```

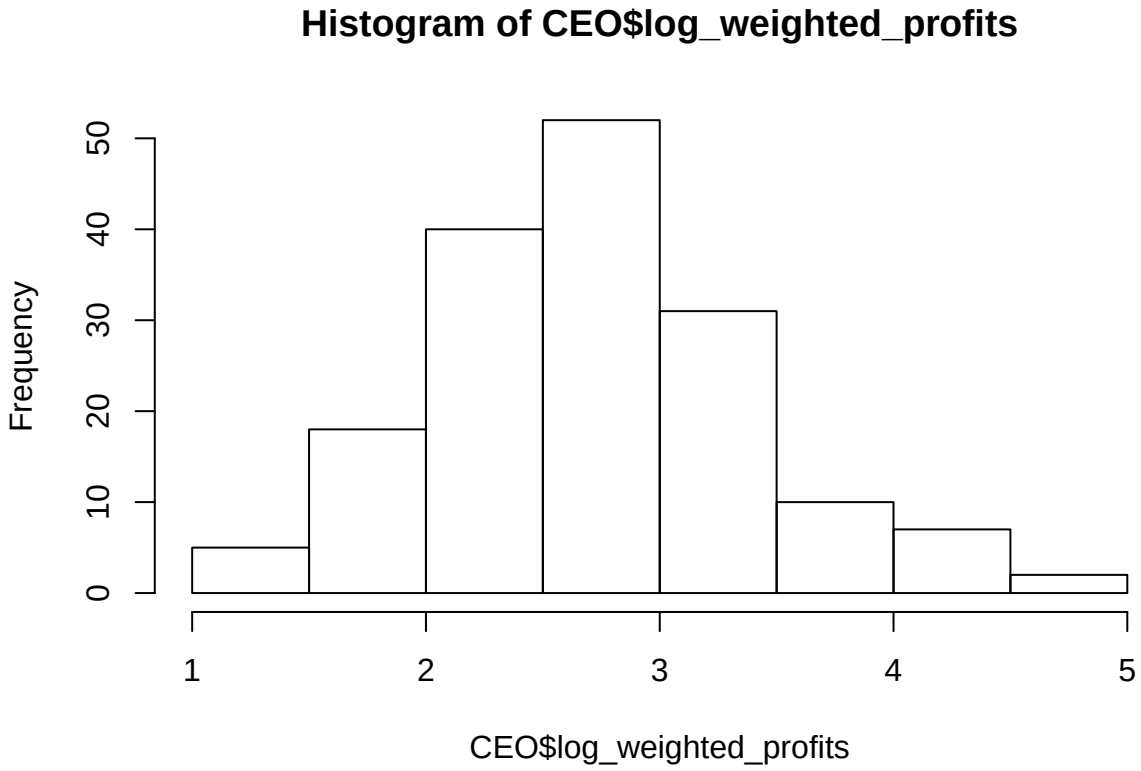
```r
cor(CEO$salary, CEO$log_weighted_mktval, use = "complete.obs")
```

```
## [1] 0.4752557
```

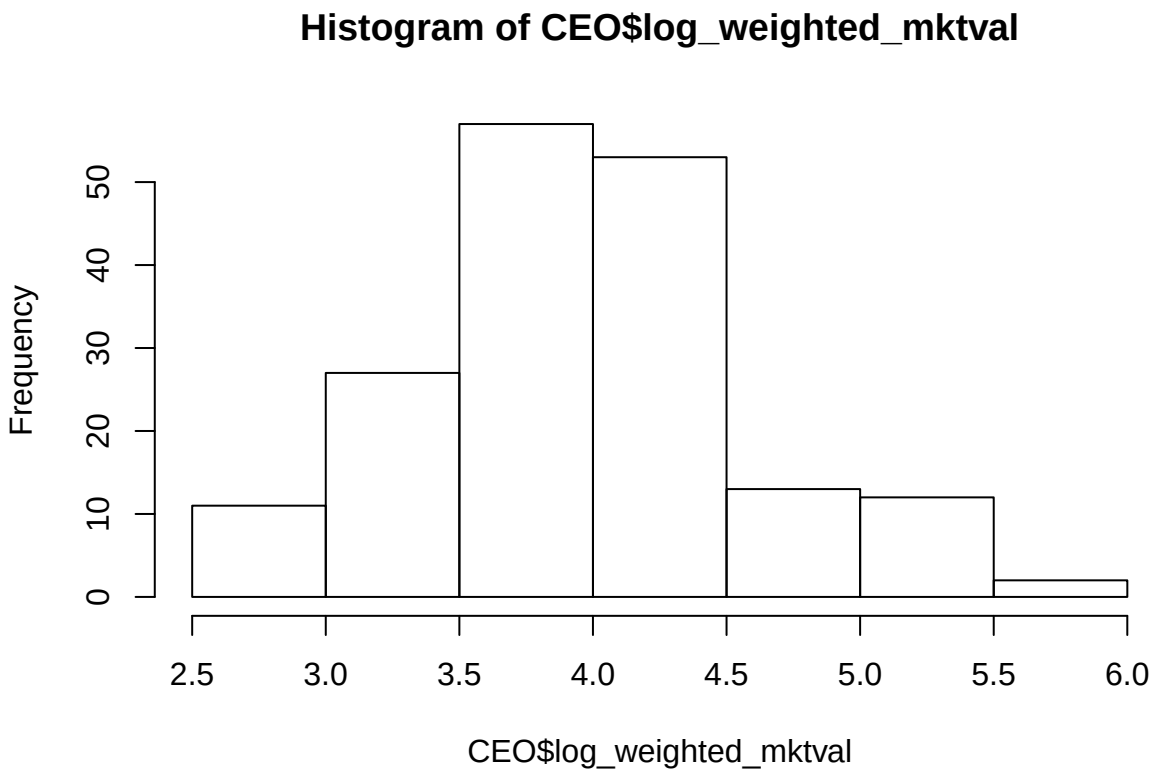
```r
cor(CEO$log_weighted_profits, CEO$log_weighted_mktval, use = "complete.obs")
```

```
## [1] 0.9107056
```

```r
hist(CEO$log_weighted_profits)
```

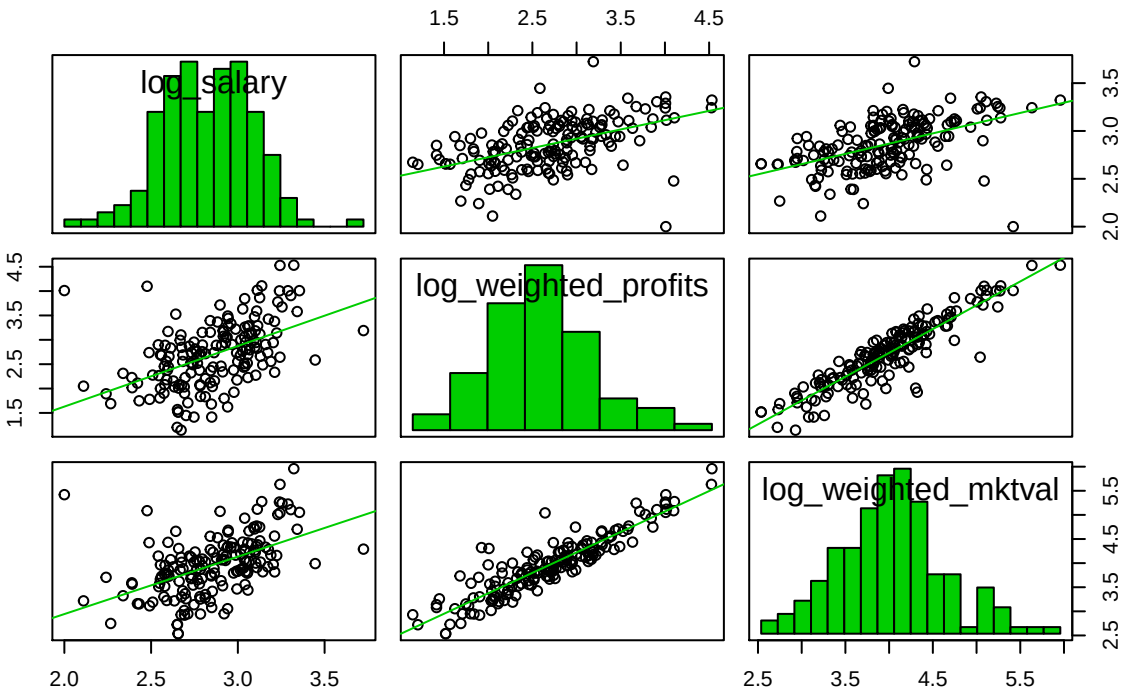



```r
hist(CEO$log_weighted_mktval)
```



Let's draw a scatterplot matrix for those weighted variables.

```
scatterplotMatrix( ~ log_salary + log_weighted_profits + log_weighted_mktval,
                   data = CEO,
#                  reg.line="" ,
                   smoother="",
                   diagonal = "histogram",
                   main = "Scatterplot Matrix for weighted profits / market value")
```

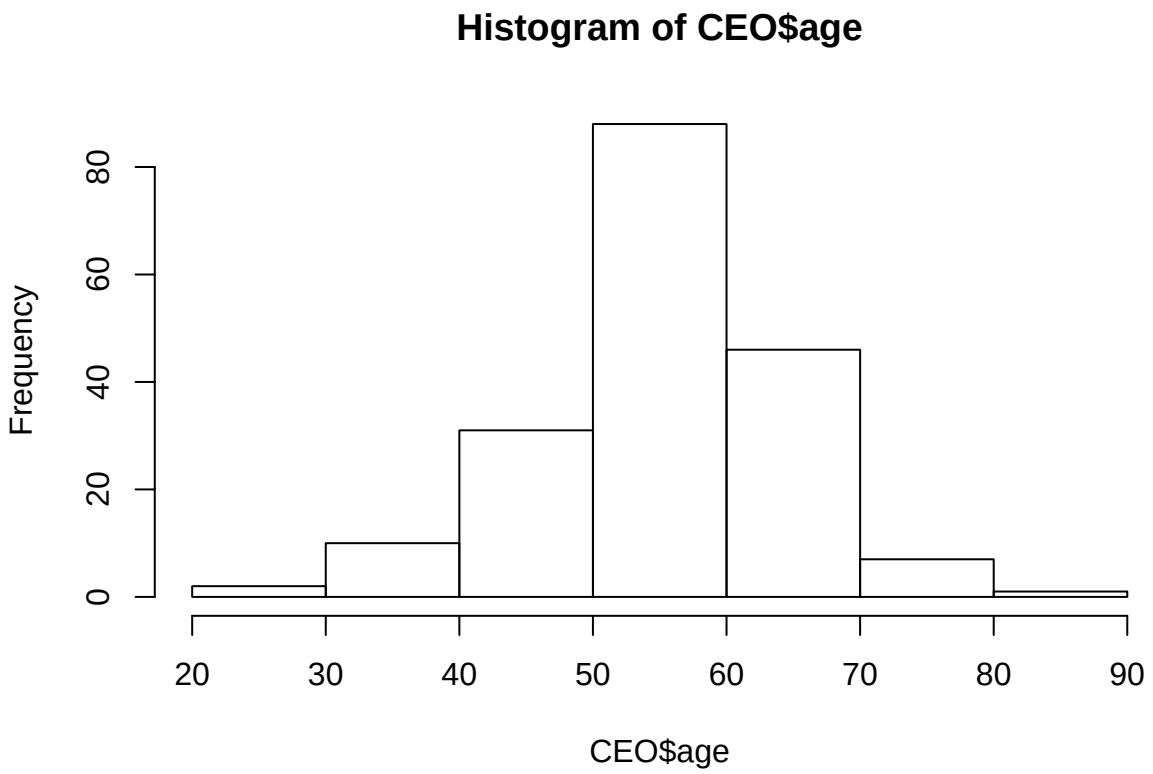


The correlation between log_salary and log_weighted_profits respectively log_salary and log_weighted_mktval don't seem stronger than their unweighted counterparts.

(However, the correlation between log_weighted_profits and log_weighted_mktval look stronger than between the unweighted versions. This possibly accounts for the intuitive assumption, that a CEOs achieved profit has a higher effect on the market value, the longer he has been CEO.)

Age factor

```
hist(CEO$age)
```



There is new significant skew in the age variable.

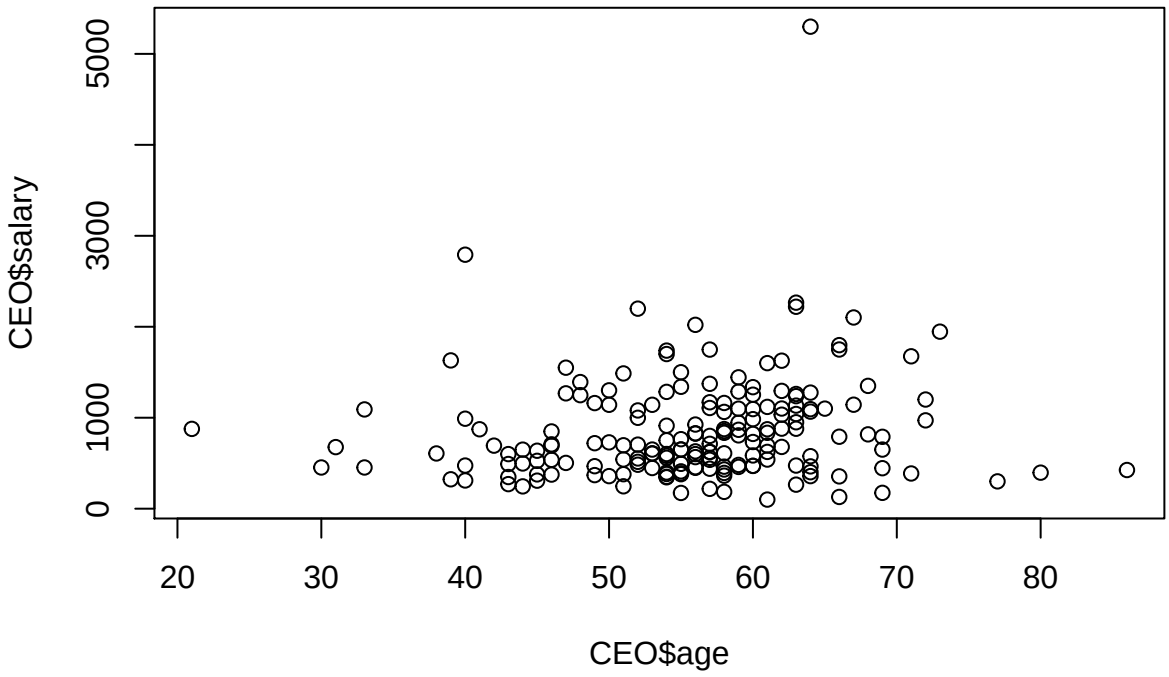
Let's check if there is a correlation between age and salary. It seems natural to assume that their might be a correlation between those two variables.

```
cor(CEO$age, CEO$salary)
```

```
## [1] 0.130081
```

There seems to be no significant linear relationship between age and salary. This can also be seen in the scatterplot.

```
plot(CEO$age, CEO$salary)
```



Education factor

Let's examine, if there is a linear correlation between salary and education.

Let's add both education variables to create one variable that indicates the CEOs overall education and check the correlation (does checking the correlation really make sense here?)

```
CEO$educ = CEO$college + CEO$grad
cor(CEO$educ, CEO$salary)
```

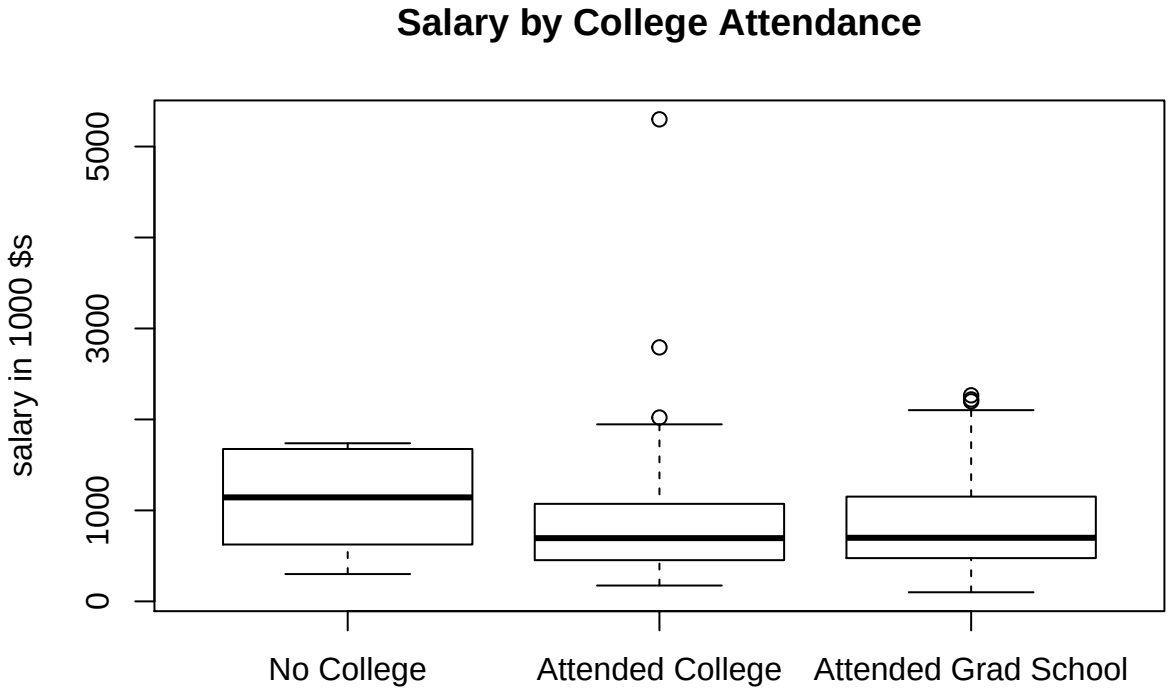
```
## [1] -0.027995
```

There seems to be no significant linear correlation. Let's look at the boxplots for all education levels.

```
educ_bin = cut(CEO$educ, breaks = 3, labels =
      c("No College", "Attended College", "Attended Grad School"))
summary(educ_bin)
```

```
##           No College     Attended College Attended Grad School
##                    5                   80                  100
```

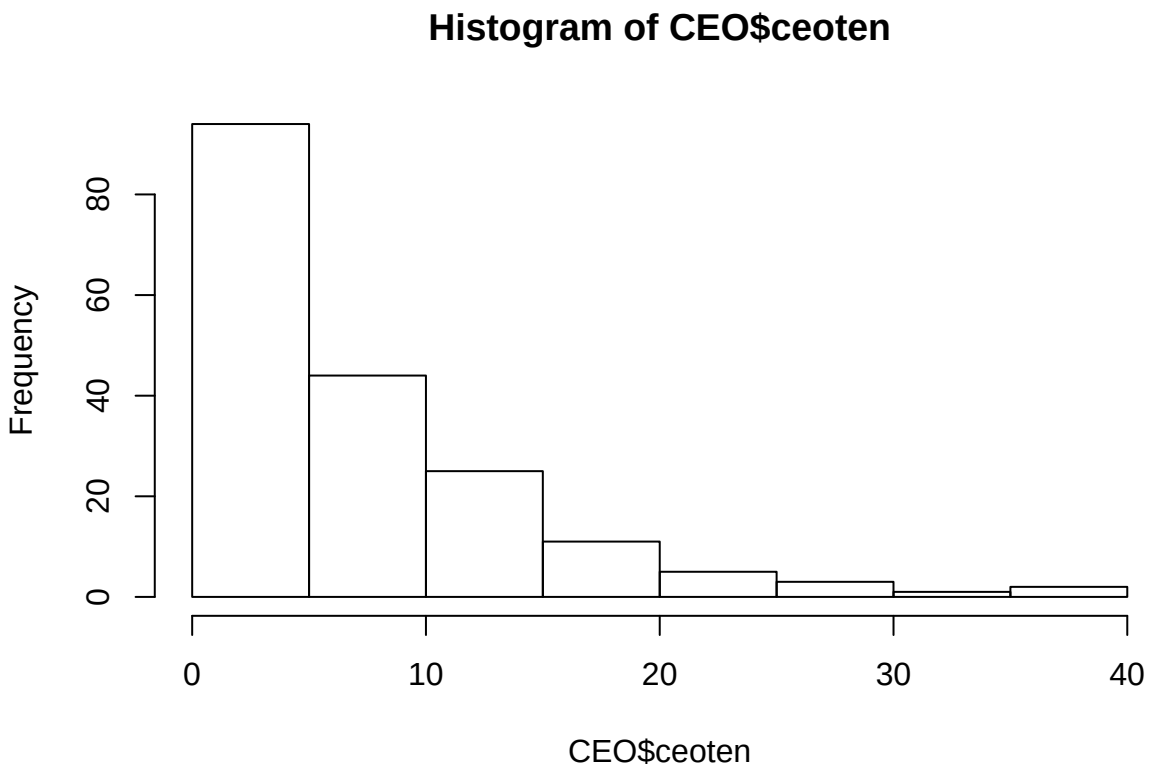
```
boxplot(salary ~ educ_bin, data = CEO,
        main = "Salary by College Attendance",
        ylab = "salary in 1000 $s")
```



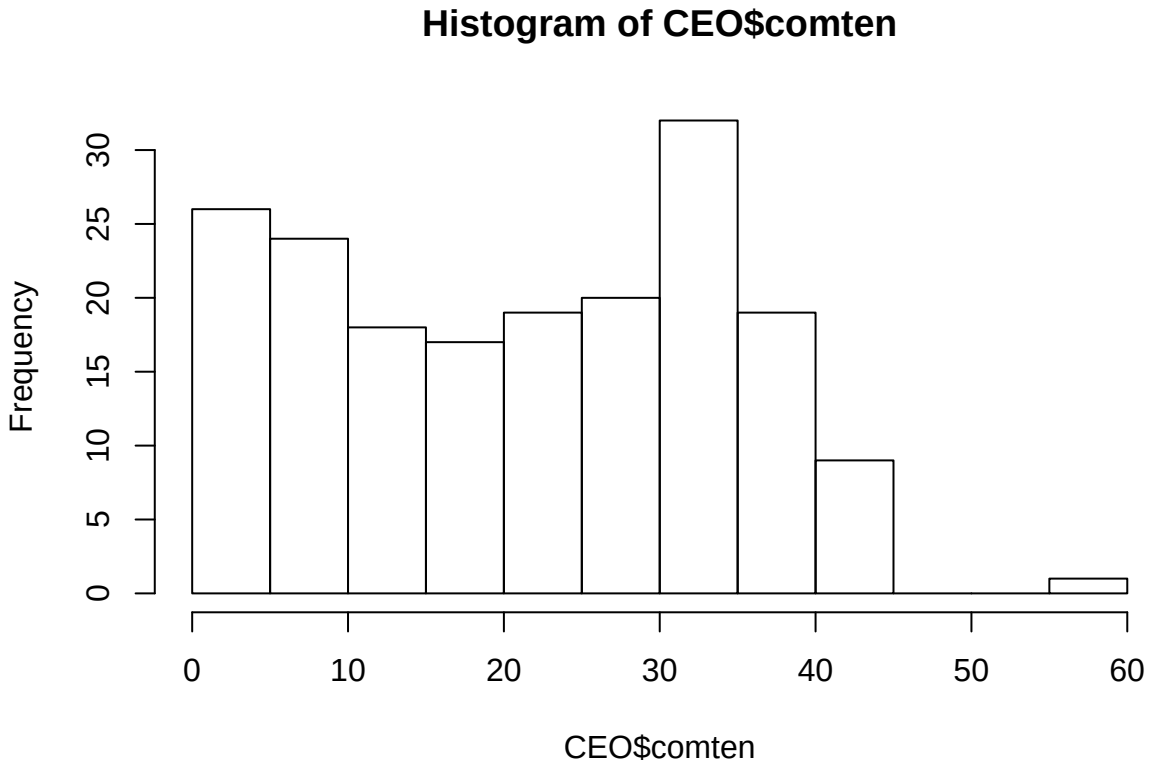
Since there are only 5 data points of CEOs with no college education, the "No College" boxplot has no significance and shall not be discussed further. The other two boxplots reveal that there is only very little difference in the salary distribution between CEOs that attended College and those that attended Grad School.

Seniority factor

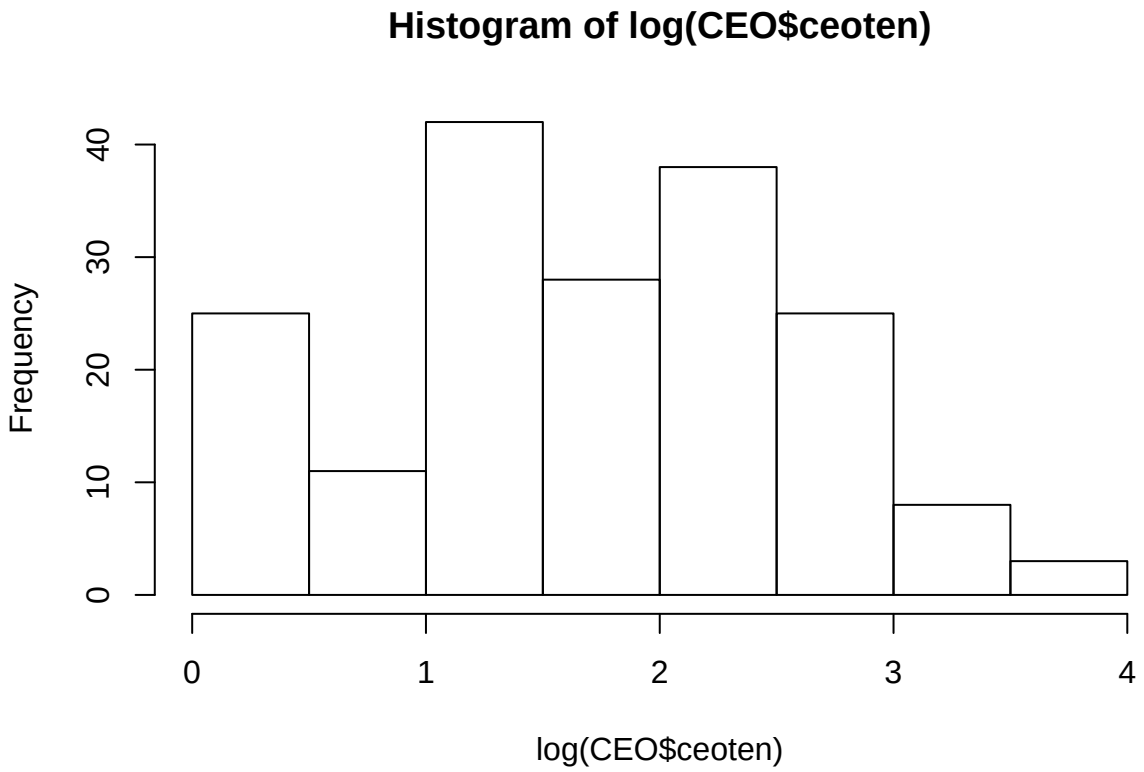
```
hist(CEO$ceoten)
```




```
hist(CEO$comten)
```




```
hist(log(CEO$ceoten))
```



The ceoten variable (amount of years the CEO has been in office within the company) is skewed. Use transformation?????

Finally, let's check out, if seniority and salary correlate in some way.

```
cor(CEO$comten, CEO$salary)
```

```
## [1] 0.06836262
```

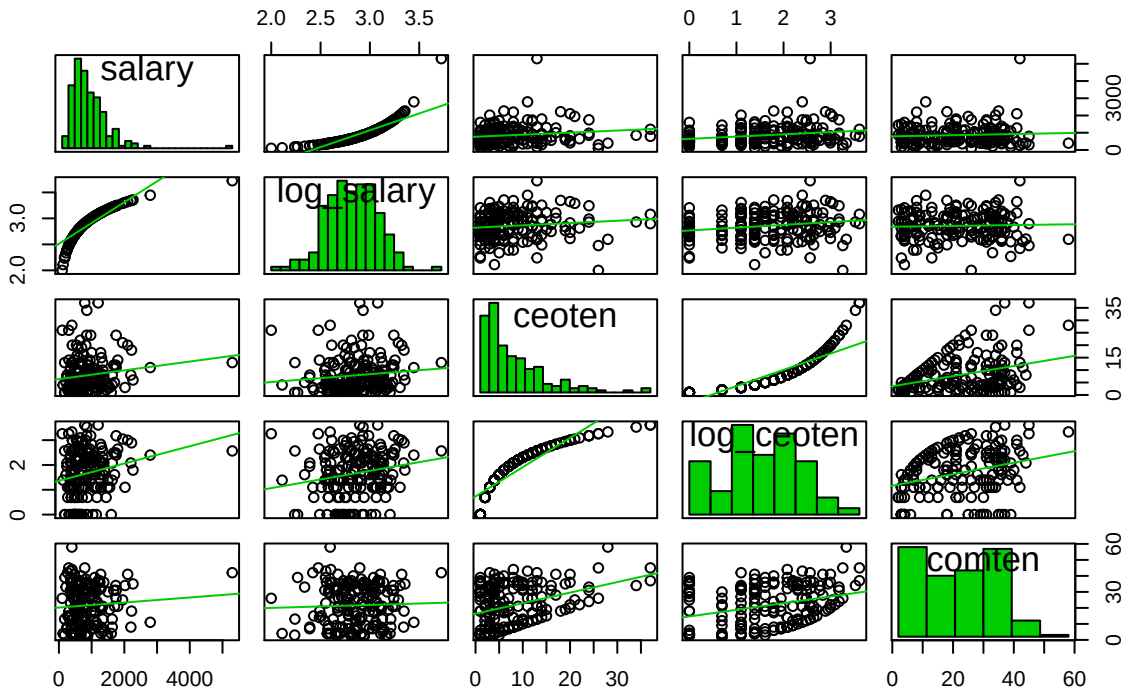
```
cor(CEO$ceoten, CEO$salary)
```

```
## [1] 0.1597714
```

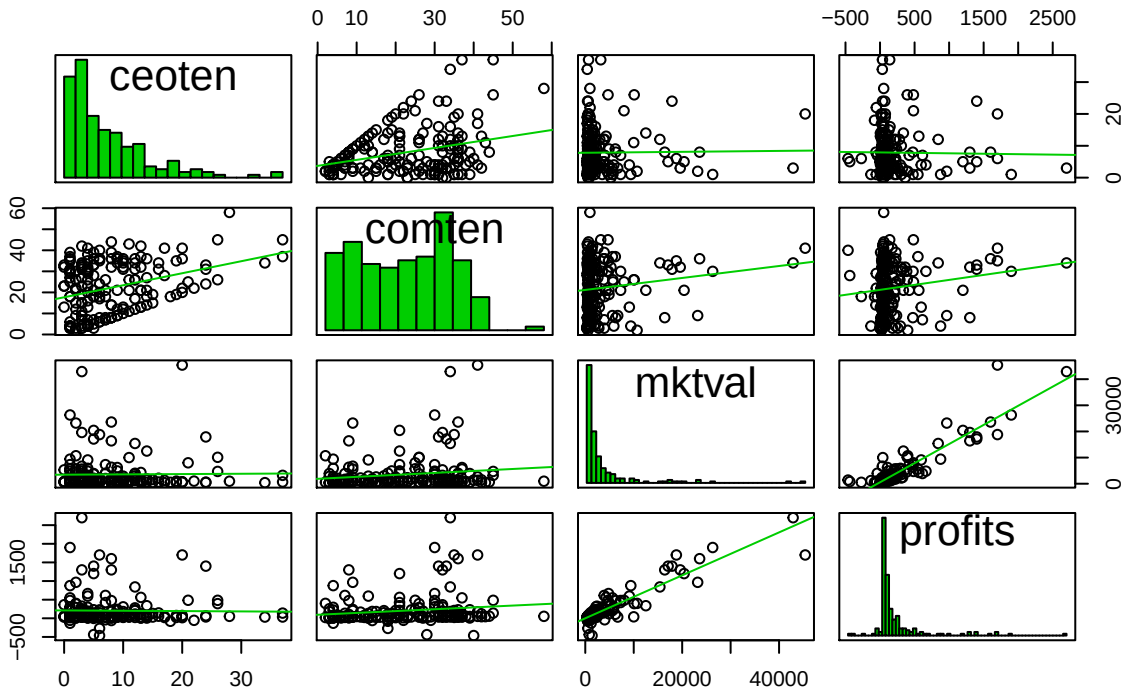
```
CEO$log_ceoten = log(CEO$ceoten)
CEO$log_ceoten[CEO$ceoten == 0] <- NA

cor(CEO$log_ceoten, CEO$log_salary, use = "complete.obs")
```

```
## [1] 0.1938044
```

```
scatterplotMatrix( ~ salary + log_salary + ceoten + log_ceoten + comten,
                   data = CEO,
#                  reg.line="" ,
                   smoother="",
                   diagonal = "histogram",
                   main = "Scatterplot Matrix Seniority")
```



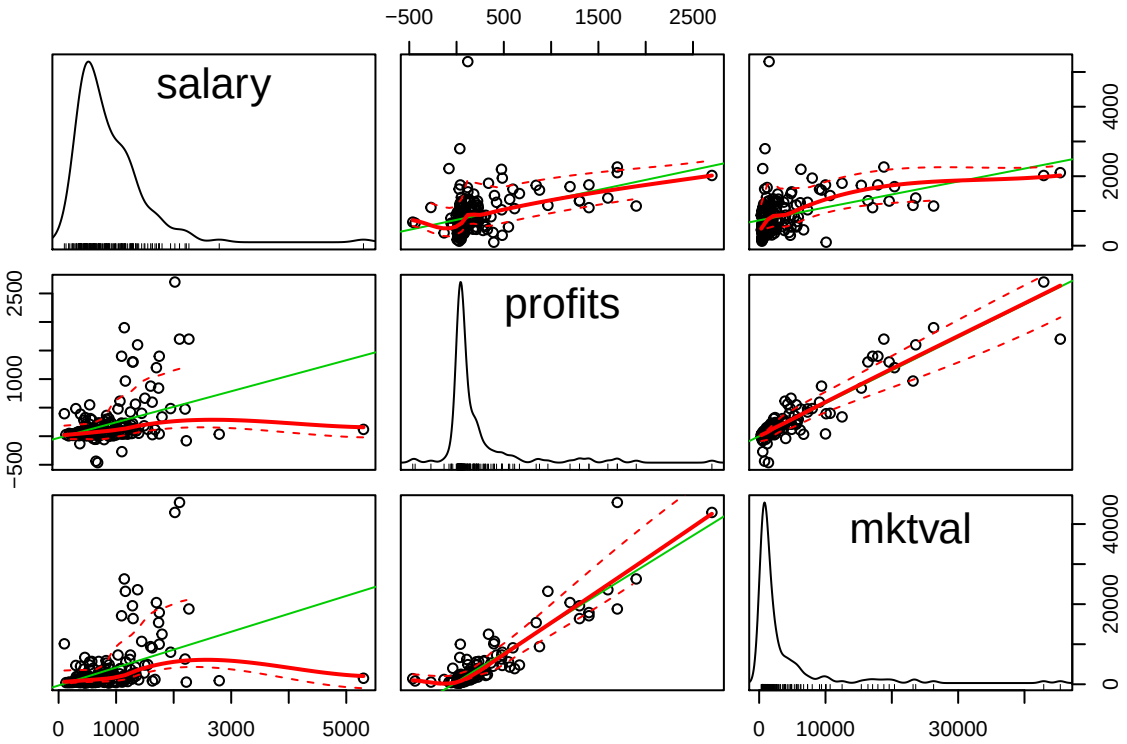

```
scatterplotMatrix( ~ ceoten + comten + mktval + profits,
                   data = CEO,
#                  reg.line="" ,
                   smoother="",
                   diagonal = "histogram",
                   main = "Scatterplot Matrix Seniority")
```



Confounding Variables

There may very well be a confounding variable effect taking place in the dataset. If we look again at the scatterplotMatrix below we can see that salary and mktval seem to be correlated. However, profits and mktval are also very highly correlated.

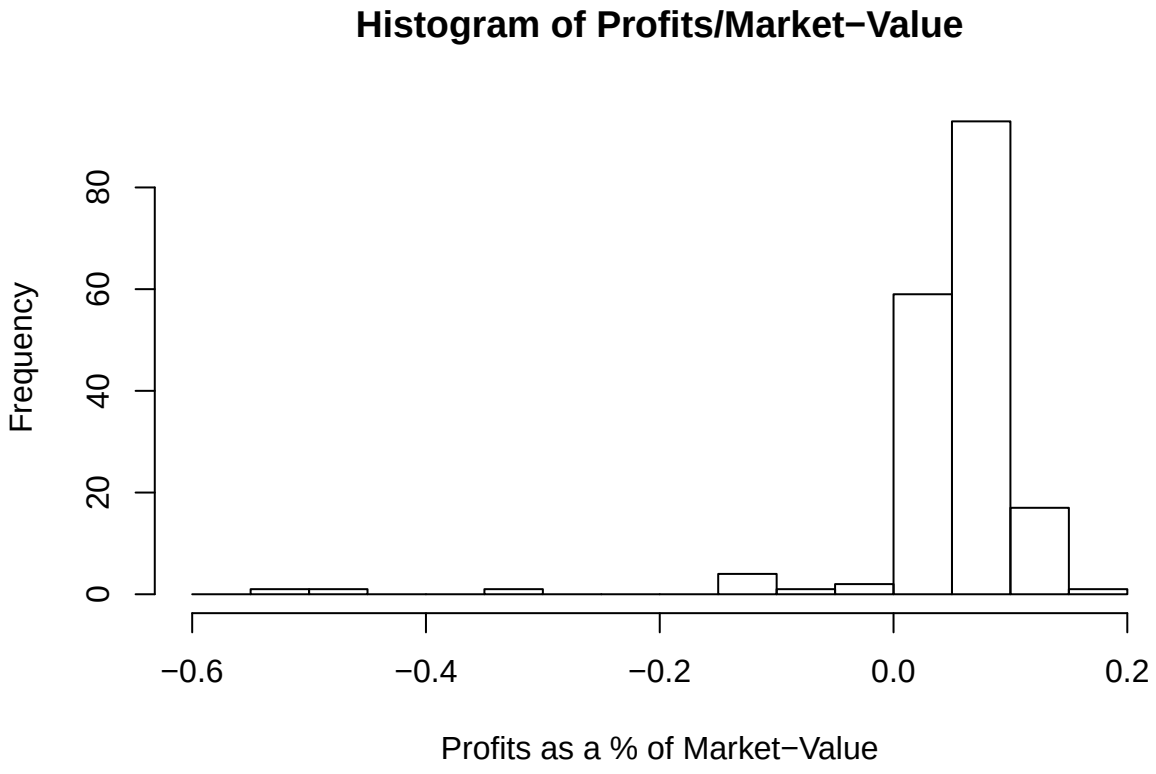
```
scatterplotMatrix(~salary + profits + mktval, data=CEO)
```



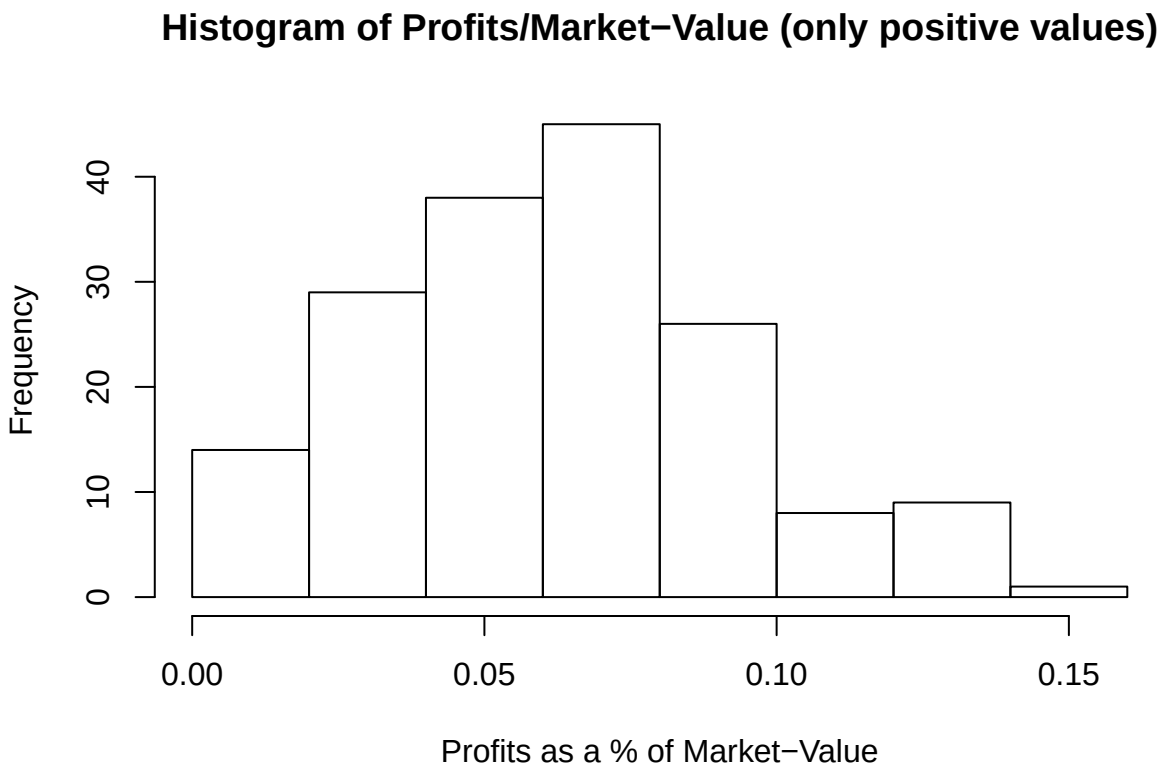
This could imply that mktval is a confounding variable. This would make intuitive sense as well because a company with a higher market value could frequently make higher profits than a company with a low market value. Similarly, a high market value company may decide to pay their CEO more money than a low market value company. In that way, marketvalue could be confounding the relationship between profits and salary and influencing a correlation which may not actually exist.

To get around this, we reduce the impact of mkval on profits by dividing profits by mktval. This should give us an intuitive variable that is the 1990 profits as a % of market value. As we can see from the histogram below, the variable is fairly normally distributed when profits are positive, with the negative profits observations being clear outliers. This can be seen more explicitly in the second histogram, where we are only looking at positive profit values.

```
CEO$prof_perc <- CEO$profits/CEO$mktval
hist(CEO$prof_perc,breaks=seq(-0.6,0.2,0.05),main="Histogram of Profits/Market-Value", xlab="Profits as
```

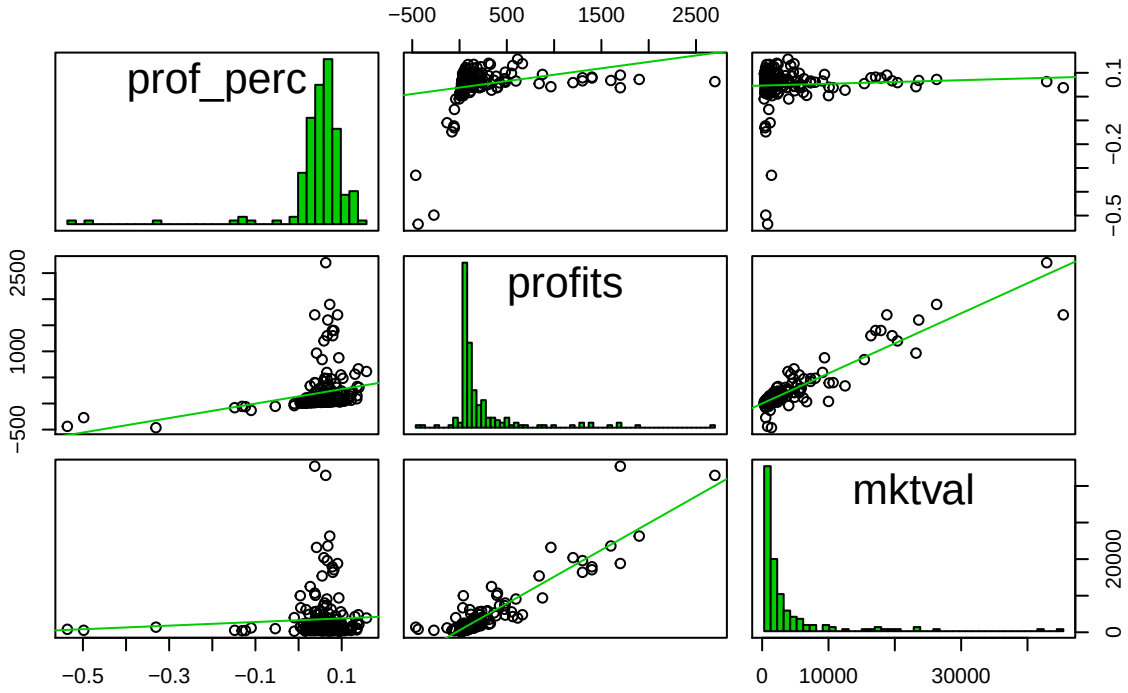



```
hist(CEO$prof_perc[CEO$profits>0],main="Histogram of Profits/Market-Value (only positive values)", xlab=
```



To ensure that the transformation has worked correctly, we can compare the new *prof_perc* variable to the *mktval* and *profits* variables.

```
# scatterpot matrix of new prof_perc variable
scatterplotMatrix(~prof_perc + profits + mktval ,data=CEO,
                  smoother="",
                  diagonal="histogram",
                  main="Scatterplot Matrix CEO Salary")
```



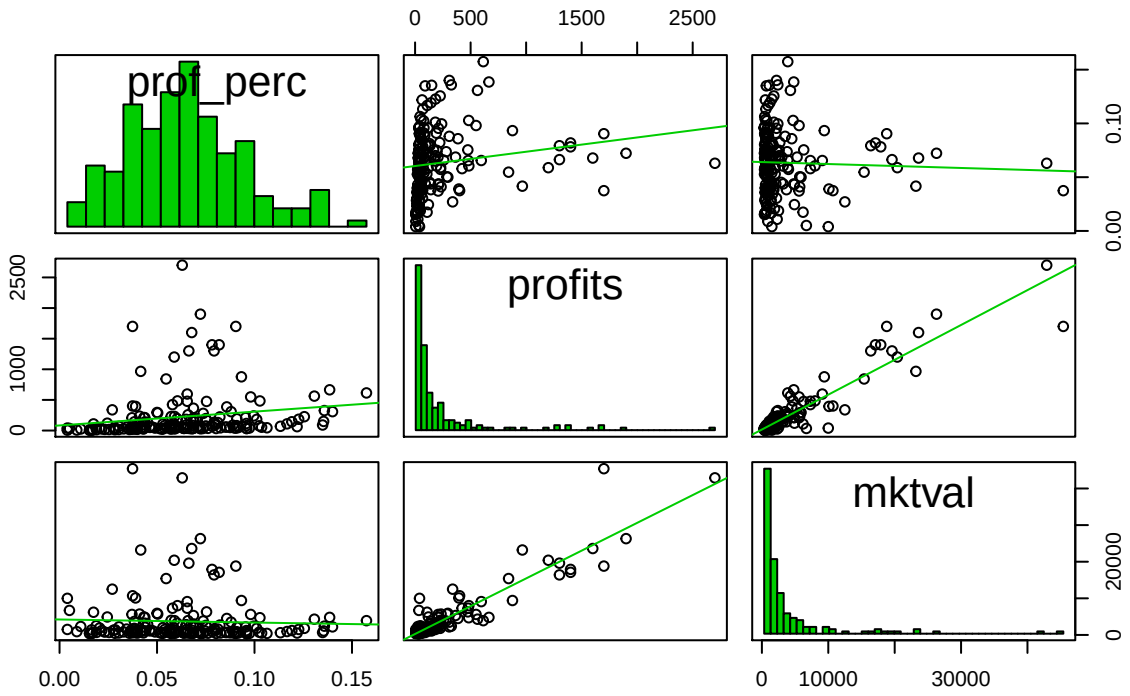

```
cor(CEO$prof_perc, CEO$mktval,use="complete.obs")
```

```
## [1] 0.06116972
```

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

```
## [1] 0.2734555
```

```
# scatterplot matrix and correlations without negative values
scatterplotMatrix(~prof_perc + profits + mktval ,data=CEO[CEO$profits>0,],
                  smoother="",
                  diagonal="histogram",
                  main="Scatterplot Matrix CEO Salary")
```




```
cor(CEO$prof_perc[CEO$profits>0], CEO$mktval[CEO$profits>0],use="complete.obs")
```

```
## [1] -0.0391175
```

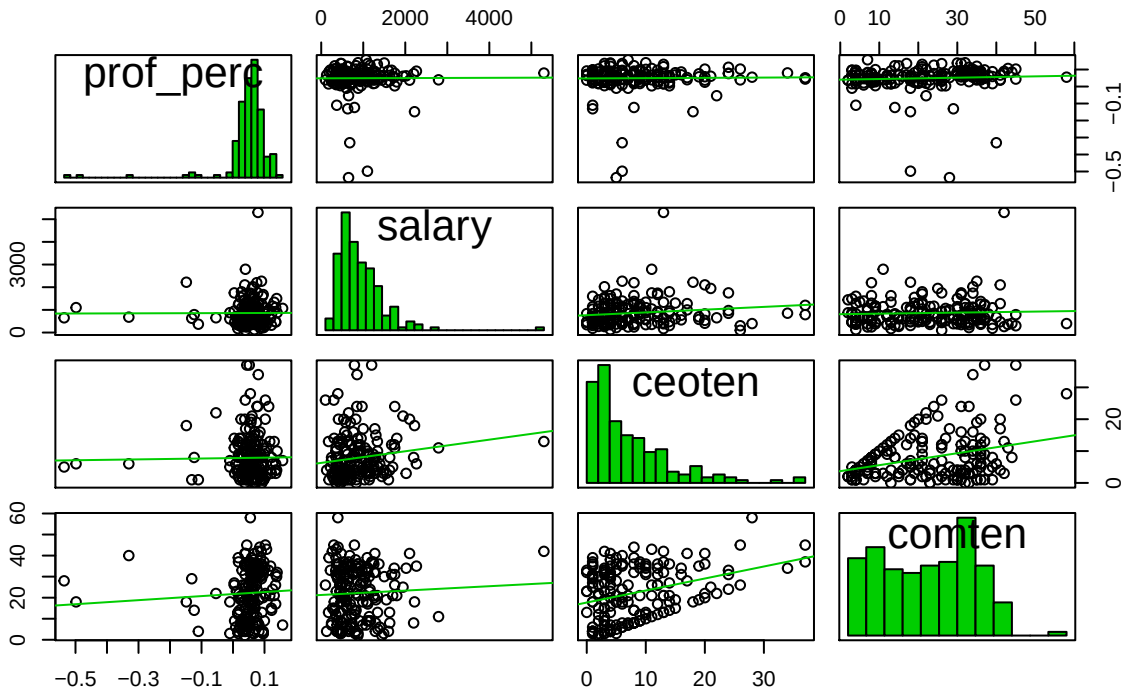
```
cor(CEO$prof_perc[CEO$profits>0], CEO$profits[CEO$profits>0],use="complete.obs")
```

```
## [1] 0.1730873
```

Finally, we can compare the new *prof_perc* variable to a few of the variables in our dataset.

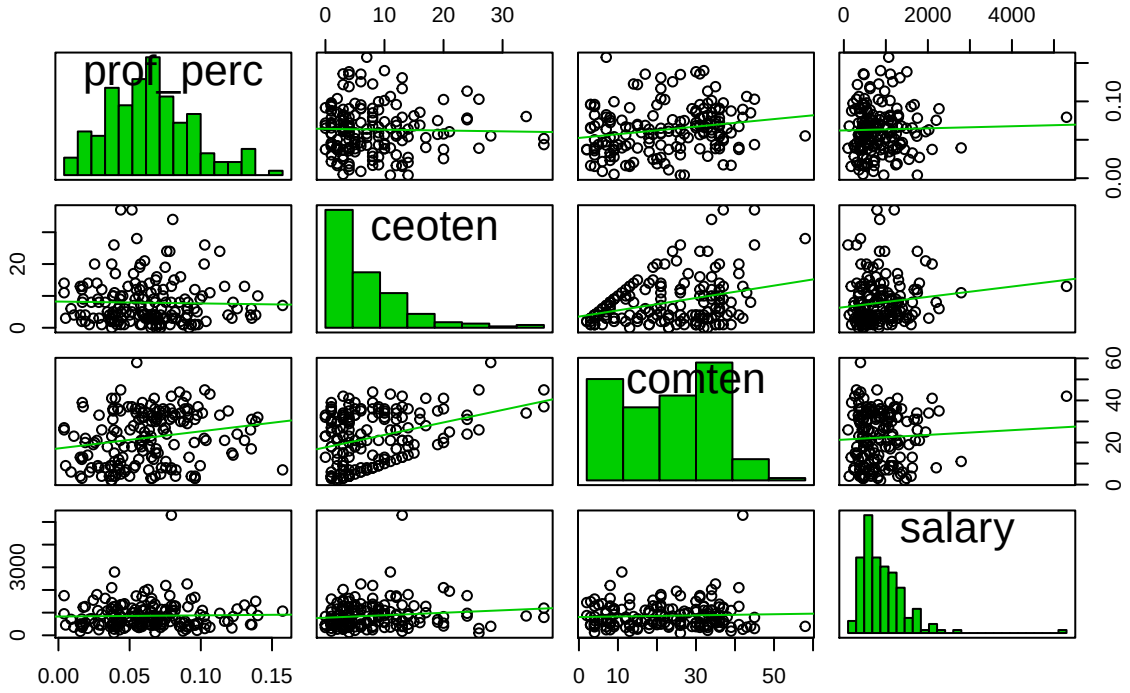
```
# prof_perc vs other variables
scatterplotMatrix(~prof_perc + salary + ceoten + comten, data=CEO,
                  smoother="",
                  diagonal="histogram",
                  main="Scatterplot Matrix CEO Salary")
```

Scatterplot Matrix CEO Salary




```
# prof_perc vs other variables with negative values removed
scatterplotMatrix(~prof_perc + ceoten + comten + salary ,data=CEO[CEO$profits>0,],
                  smoother="",
                  diagonal="histogram",
                  main="Scatterplot Matrix CEO Salary (Negative Profits Removed)")
```

Scatterplot Matrix CEO Salary (Negative Profits Removed)



```r
# correlation with and without negative values
cor(CEO$prof_perc, CEO$salary, use="complete.obs")
```

```
## [1] 0.005186691
```

```r
cor(CEO$prof_perc[CEO$profits>0], CEO$salary[CEO$profits>0], use="complete.obs")
```

```
## [1] 0.02598916
```

The new variable *prof_perc* has little to no correlation with the salary variable. This could indicate, that once the effects of market-value are removed there is very little correlation between profit and salary. This would need to be addressed further in a more rigorous approach.