Lab 1: Exploratory Analysis of CEO Salary Data

Carmen Easterwood & Andrew Kabatznick

January 29, 2016

Introduction

This report will analyze several factors that affect CEO salaries, and in particular the relationship between CEO salary and company performance. We have collected data on a sample of 185 CEOs in 1990, and will perform an exploratory analysis of this data in the sections that follow. However, we will also note any features that would be relevant to statistical modeling.

Since our primary goal is to determine the relationship between CEO salary and company performance, we have divided our variables into the following categories:

- Outcome (dependent) variable: Salary
- Key predictor (independent) variables: Profits, Mktval
- Secondary predictor (independent) variables: Age, Education Level (college & grad), Tenure (ceoten & comten)

As you read this report, keep the following caveats in mind:

1. This analysis is not causal. We will be able to show a positive relationship between CEO salary and company performance, but we are not able to say that good company performance *causes* high CEO salaries. Causation could go in the opposite direction, or there could be some third factor that drives both CEO salary and company performance (for example, an economic boom).
2. Our Salary variable is a measure of direct compensation through salary and bonuses. We do not have data on other forms of compensation, such as stock options.
3. There are some additional variables that may affect CEO salary, that we are unable to account for. These include, but are not limited to, industry, location, gender, and race.

```
setwd("~/Desktop/MIDS/Statistics/stats_lab1")
ceosal <- load("ceo_w203.RData", ceo.env <- new.env())
ceo.df <- ceo.env[["CEO"]]
```

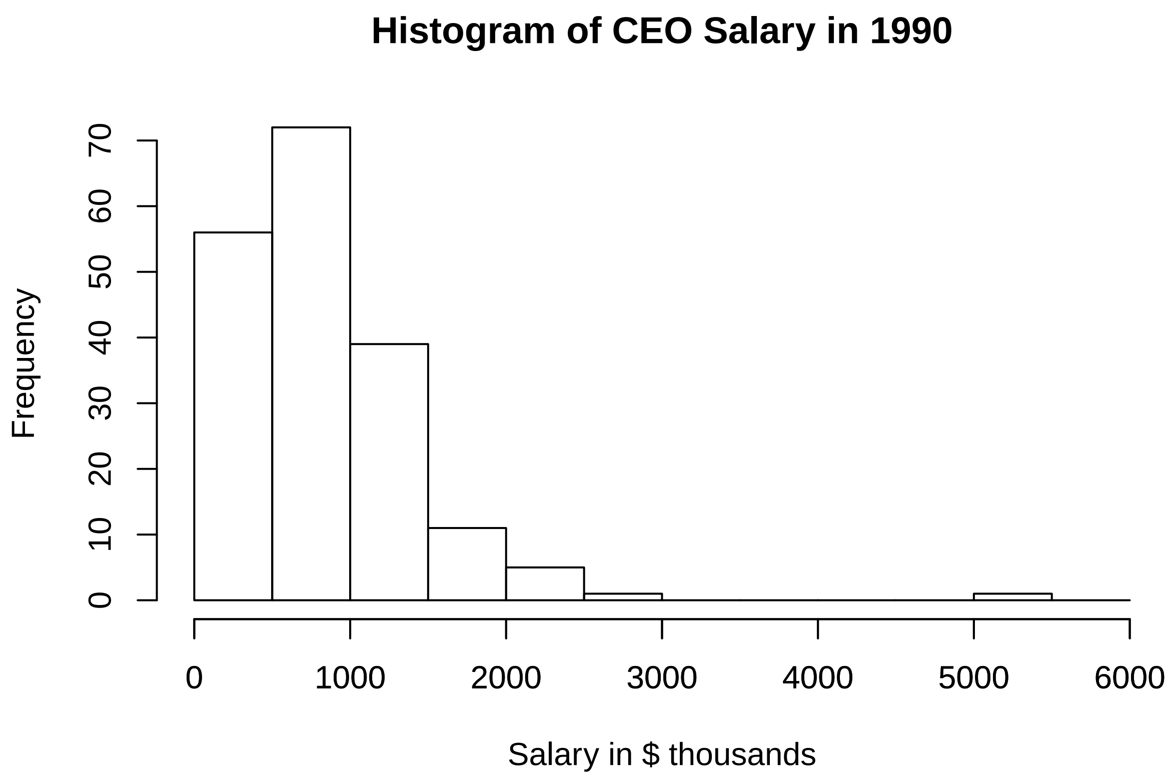
Univariate Analysis of Key Variables

In this section we will analyze each variable individually and note some of their key features.

Salary

CEO salary distribution is strongly skewed right.

```
hist(ceo.df$salary, main = "Histogram of CEO Salary in 1990",
     xlab = "Salary in $ thousands", breaks = seq(0, 6000, by = 500))
axis(1, at = seq(0, 6000, by = 1000))
```



Median salary is $697 thousand, and there is one extreme outlier at $5.3 million.

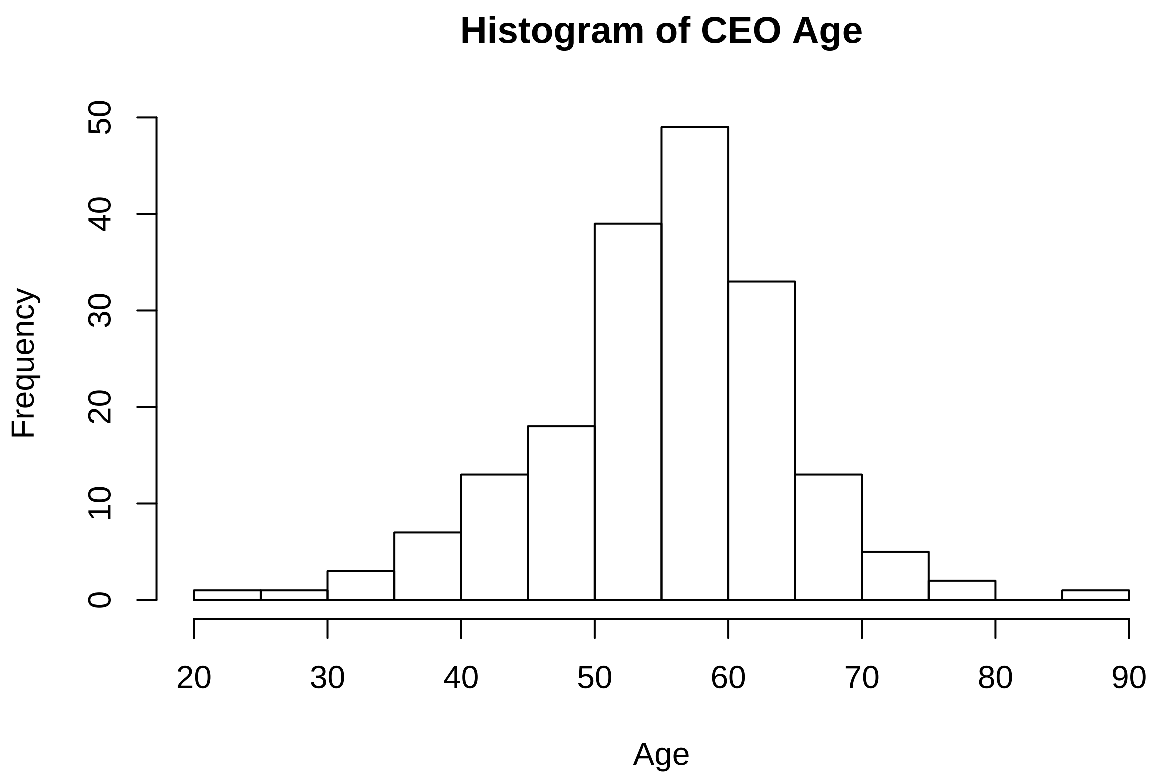
```
summary(ceo.df$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   467.0   697.0   852.9  1101.0  5299.0
```

Age

CEO age peaks between 50 and 65 years old, but ranges all the way from 21 to 86.

```
hist(ceo.df$age, breaks = 14, main = "Histogram of CEO Age", xlab = "Age")
```




```
summary(ceo.df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   51.00   57.00   55.78   61.00   86.00
```

The variance and standard deviation of Age are large:

```
var(ceo.df$age)
```

```
## [1] 85.37996
```

```
sd(ceo.df$age)
```

```
## [1] 9.240128
```

College Degree

College is a dummy variable that takes a value of 1 if the CEO is a college graduate and 0 otherwise.

```
pct.college <- (sum(ceo.df$college) / length(ceo.df$college))
```

96.2% of the CEOs in this dataset are college graduates.

Graduate Degree

Grad is a dummy variable that takes a value of 1 if the CEO holds an advanced degree and 0 otherwise.

```
pct.grad <- (sum(ceo.df$grad) / length(ceo.df$grad))
```

55.1% of the CEOs in this dataset hold advanced degrees.

Years With Company

The Median (21) and Mean (21.7) both indicate that in general CEOs have been with their companies around 21 years. However, the data is highly variable with a minimum of 2 and a maximum of 58 years.

```
summary(ceo.df$comten)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   21.00   21.66   33.00   58.00
```

Variance of Years with Company

```
var(ceo.df$comten)
```

```
## [1] 160.2132
```

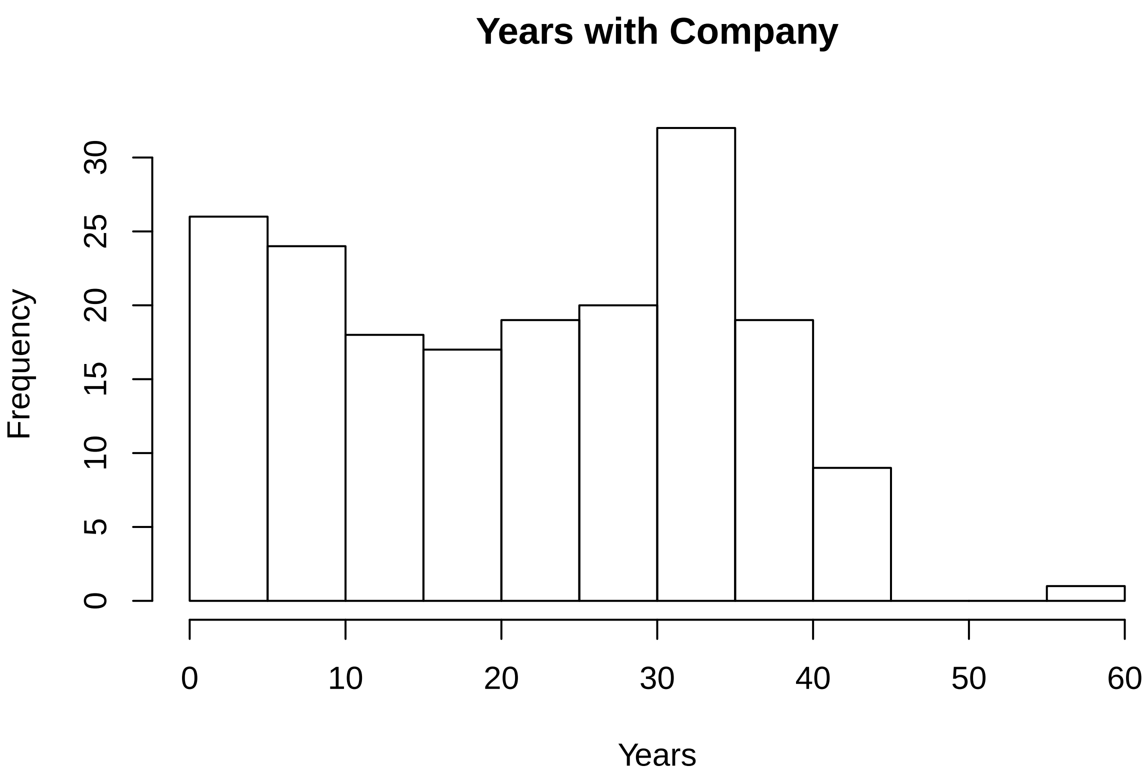
The standard deviation for years with the company is ~12.7, which is large but unsurprising given the variablity of this variable.

```
sd(ceo.df$comten)
```

```
## [1] 12.65753
```

Histogram of Years with Company

```
hist(ceo.df$comten, main="Years with Company", xlab = "Years")
```



Years as CEO

The years as CEO variable exhibits a large right skew with a median of 5 years. This skew is not entirely suprising given CEOs tend to be established in their careers and thus older, which may explain while things tail off. We must also wonder if there is a large turnover in the first few years of a CEOs tenure since half of CEOs have spent 0-5 years in their position.

Summary Statistics for Years as CEO

```
summary(ceo.df$ceoten)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   5.000   7.681  11.000  37.000
```

Variance of Years as CEO

```
var(ceo.df$ceoten)
```

```
## [1] 50.65317
```

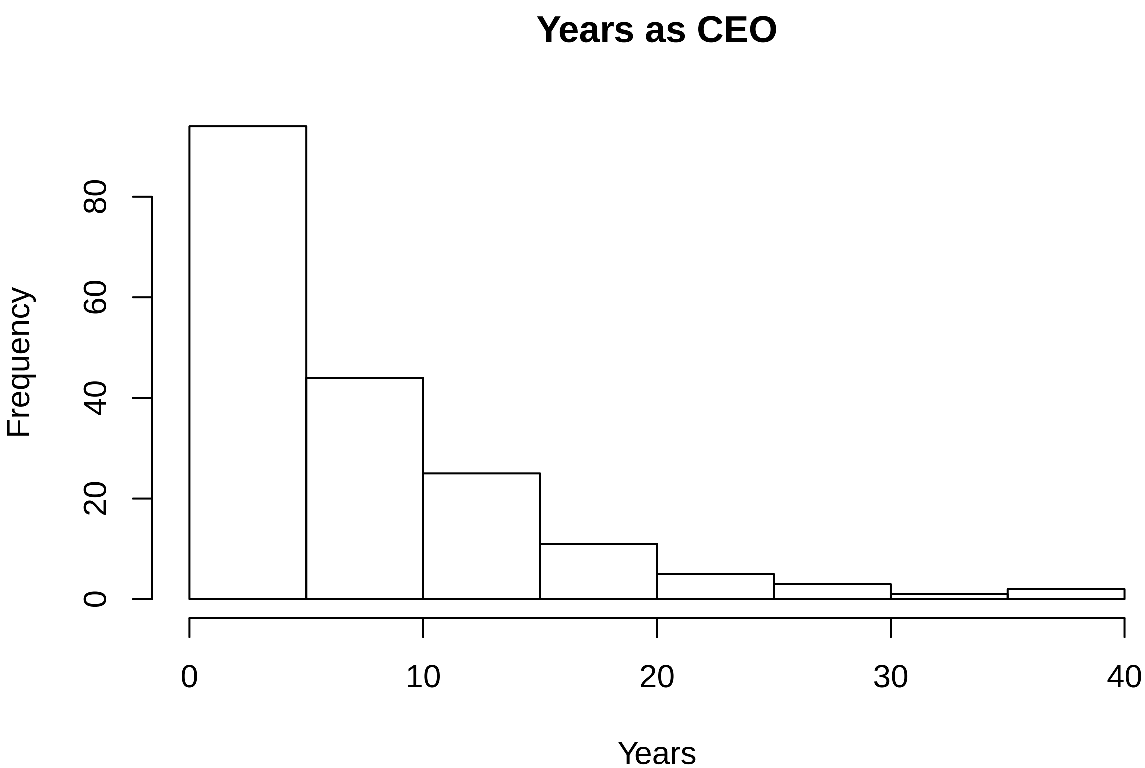
Standard Deviation of Years as CEO

```
sd(ceo.df$ceoten)
```

```
## [1] 7.117104
```

Histogram of Years as CEO

```
hist(ceo.df$ceoten, main="Years as CEO", xlab = "Years")
```



Profits

Companies in the sample tend to make a profit as shown by the median profit of $57 million. The mean is significantly higher than the median, driven by a notable outlier of $2.7 billion. While profits can be either positibe or negative, we found 5 datapoints that are equal to negtive 1. Given that CEOs with this profit value also have a market value of -1, we believe this is a bad value.
Does "bad value" mean poor company performance or a coding error? Same question for the market value section

Summary Statistics for Profits

```r
summary(ceo.df$profits)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -463.0    33.0    57.0   199.2   195.0  2700.0
```

Variance of Profits

```r
var(ceo.df$profits)
```

```
## [1] 158154.2
```

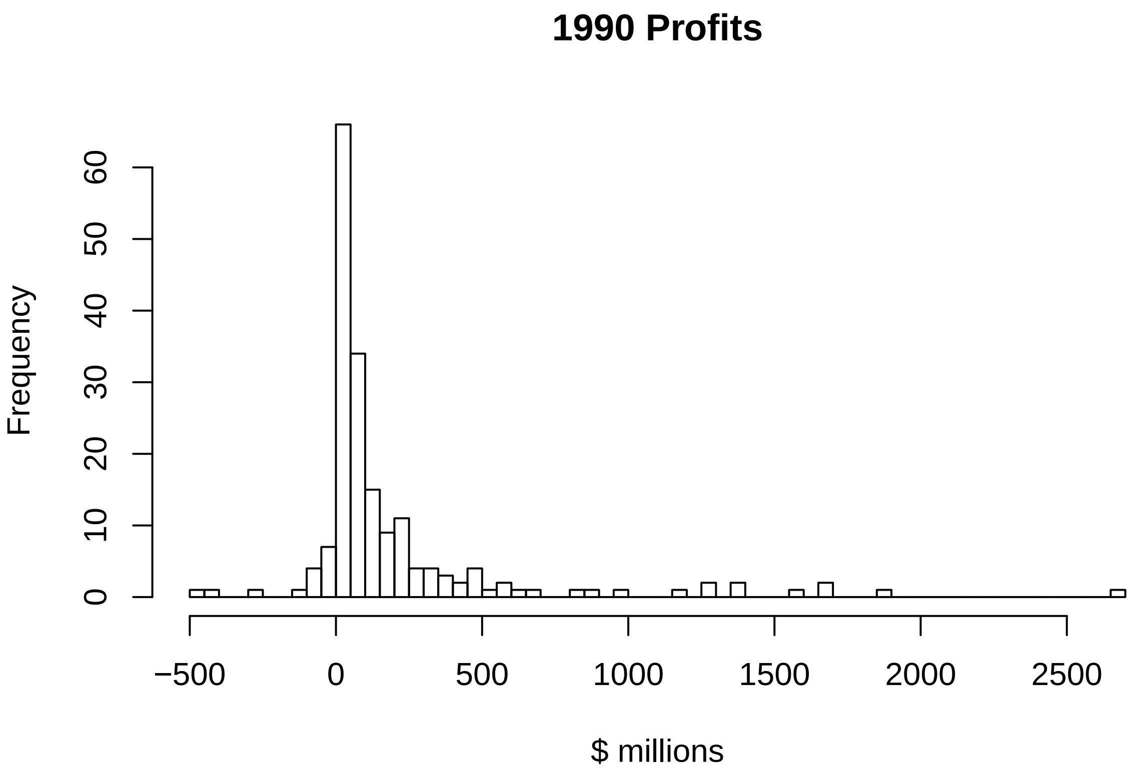
Standard Deviation of Profits

```r
sd(ceo.df$profits)
```

```
## [1] 397.686
```

Histogram of Profits

```r
hist(ceo.df$profits, main="1990 Profits",
     xlab = "$ millions", breaks = 100)
```



Market Value

Unsurprisingly, the Market Value for companies has a right skew, similar to Profits. While the median value is $1.2 billion, there are 2 outliers above $40 billion. The minimum of -1 looks to be an error, as Market Values should range from 0 to infinity. We found 5 datapoints that are equal to negative 1, which appears to be a bad value. *Does "bad value" mean poor company performance or a coding error? Same question for the profits section*

Summary Statistics for Market Value

```r
summary(ceo.df$mktval)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      -1     567    1200    3450    3200   45400
```

Variance of Market Value

```r
var(ceo.df$mktval)
```

```
## [1] 40202491
```

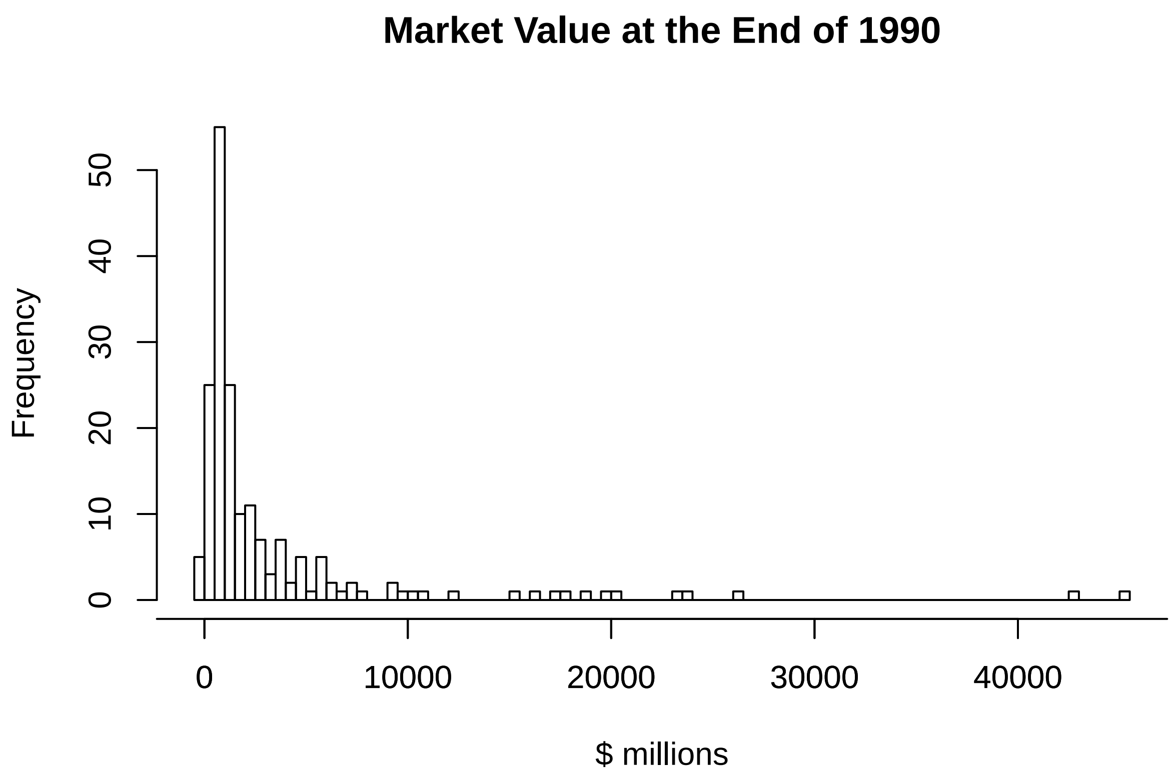
Standard Deviation of Market Value

```r
sd(ceo.df$mktval)
```

```
## [1] 6340.543
```

Histogram of Market Value

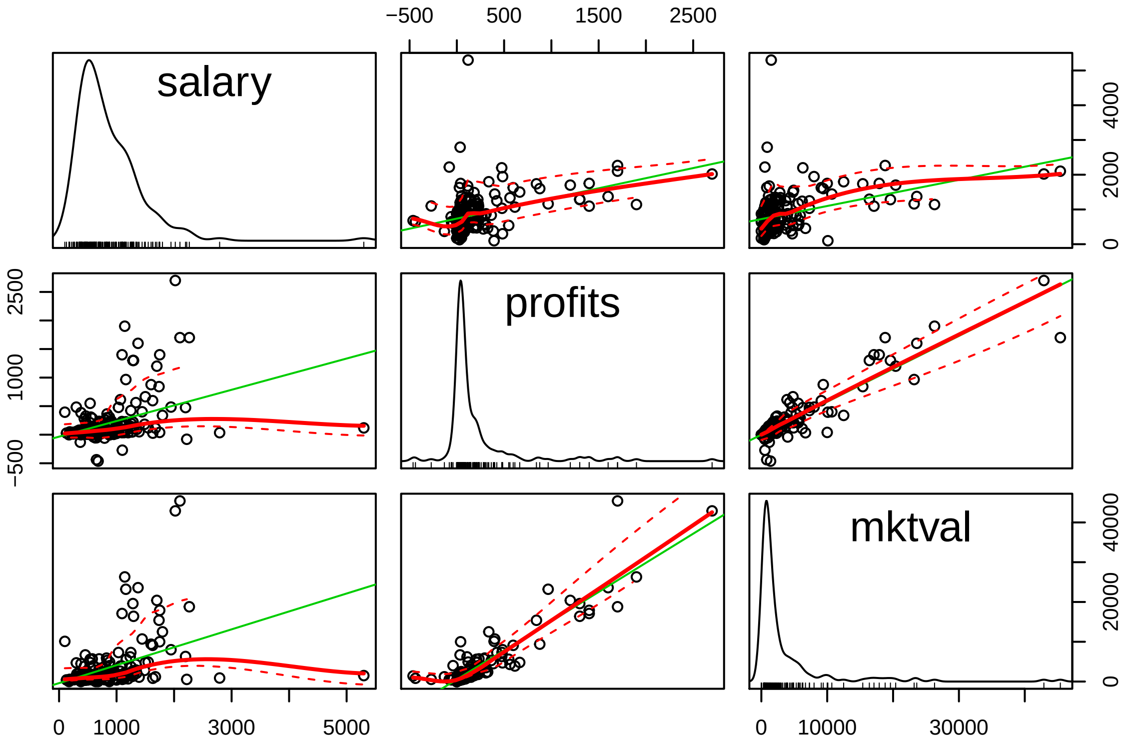
```r
hist(ceo.df$mktval, main = "Market Value at the End of 1990",
     xlab = "$ millions", breaks = 100)
axis(1, at = seq(-10000, 50000, by = 10000))
```



Key Bivariate Relationships

In this section we analyze some key bivariate relationships. Recall from the introduction that we defined our key variables are Salary, Profits, and Market Value. Below is a scatterplot matrix of these key variables.

```r
library(car)
scatterplotMatrix(~ salary + profits + mktval, data = ceo.df)
```



Profits & Market Value

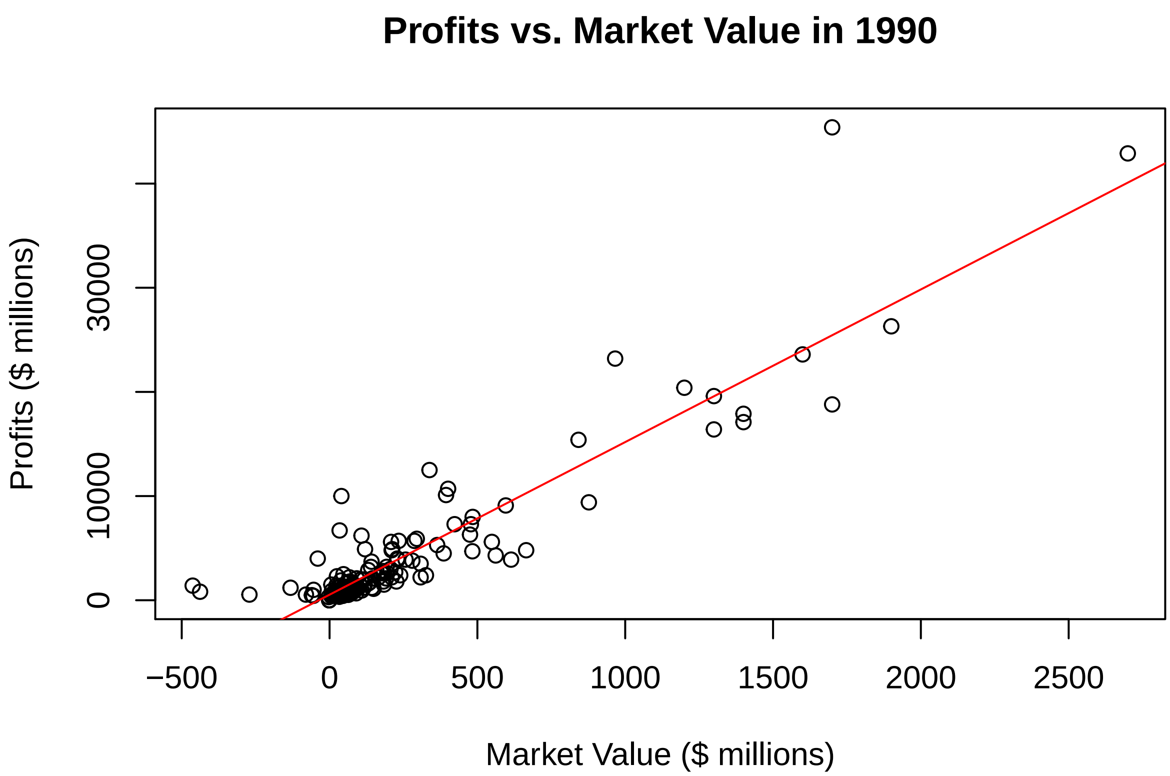
These are our two measures of company performance, so we confirm they are positively correlated.

```
cor(ceo.df$profits, ceo.df$mktval)
```

```
## [1] 0.9190233
```

For the most part, as profits increase, so does market value. The relationship appears to hold as long as profits are nonnegative, since market value cannot go below zero.

```
plot(ceo.df$profits, ceo.df$mktval,
     main = "Profits vs. Market Value in 1990",
     xlab = "Market Value ($ millions)",
     ylab = "Profits ($ millions)")
abline(lm(ceo.df$mktval ~ ceo.df$profits), col = "red")
```



Profits & CEO Salary

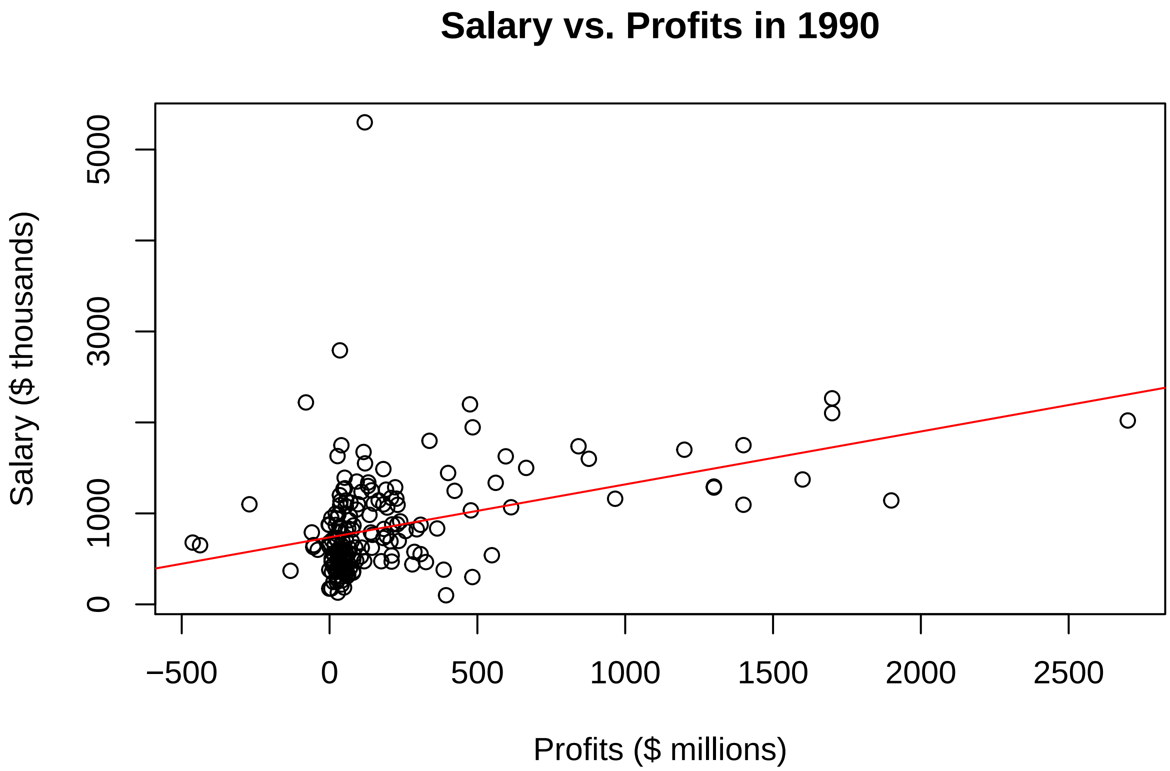
Salary and Profits don't have a strong linear relationship as demonstrated by the low correlation.

```
cor(ceo.df$salary, ceo.df$profits)
```

```
## [1] 0.3989609
```

When profits are between -$500 million and $500 million and salary is less than $2 million, there seems to be a slight linear relationship. However, outlier CEOs, who significantly outperformed their peers by making significant profits per dollar earned, skew the linearity of the data.

```
plot(ceo.df$profits, ceo.df$salary,
     main = "Salary vs. Profits in 1990",
     xlab = "Profits ($ millions)",
     ylab = "Salary ($ thousands)")
abline(lm(ceo.df$salary ~ ceo.df$profits), col = "red")
```



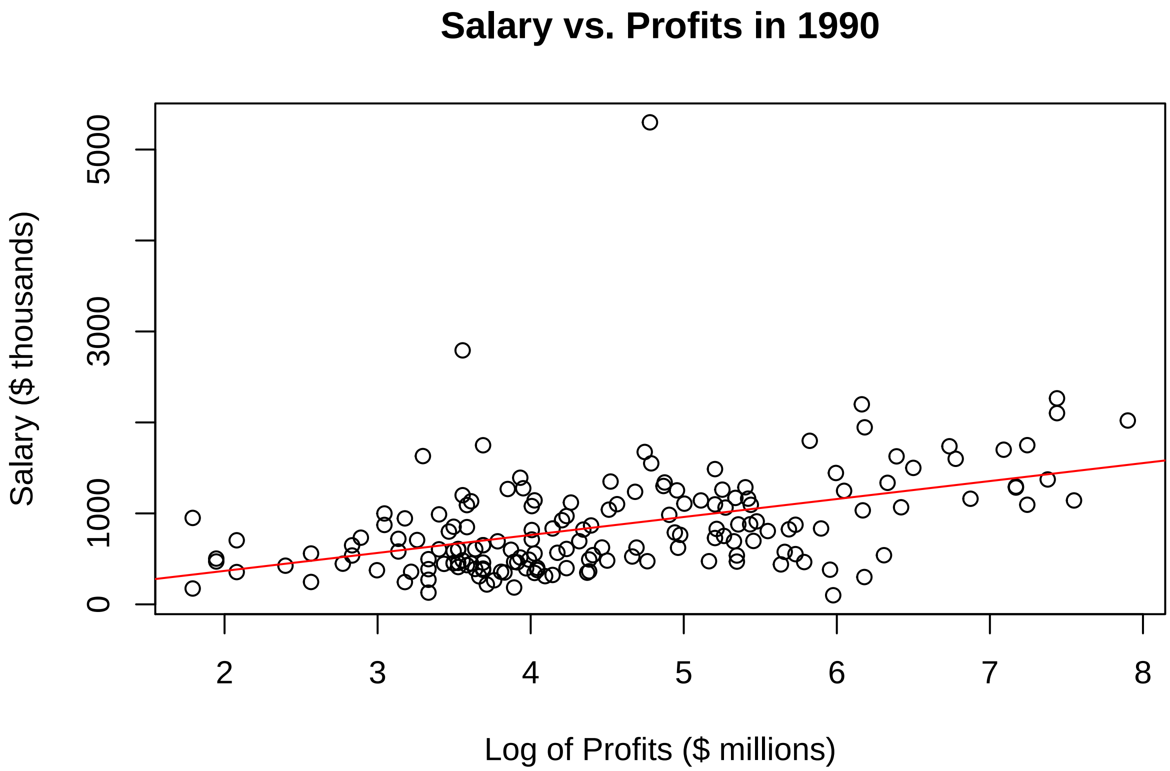
However, after performing a logorithmic transformation on the profits variable, the variables show more a linear relationship.

```
plot(log(ceo.df$profits), ceo.df$salary,
     main = "Salary vs. Profits in 1990",
     xlab = "Log of Profits ($ millions)",
     ylab = "Salary ($ thousands)")
```

```
## Warning in log(ceo.df$profits): NaNs produced
```

```
abline(lm(ceo.df$salary ~ log(ceo.df$profits)), col = "red")
```

```
## Warning in log(ceo.df$profits): NaNs produced
```



Market Value & CEO Salary

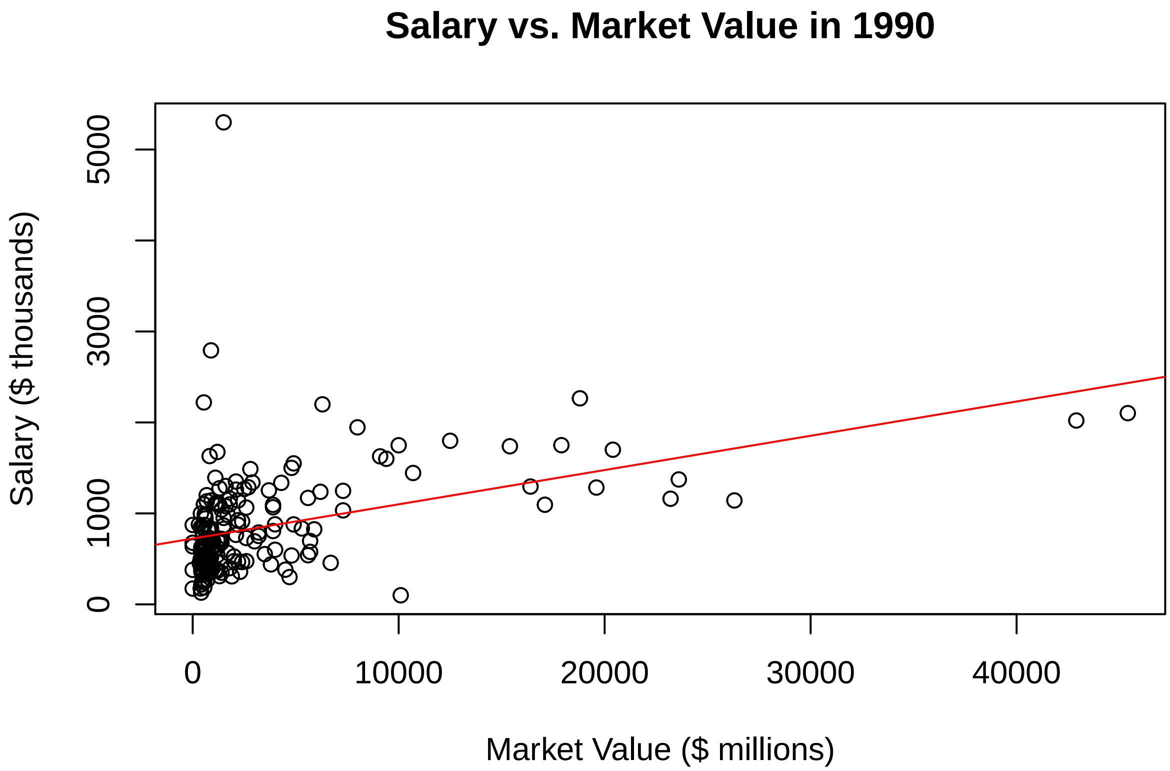
Like the relationship between salary and profits, salary and market value also don't have a strong linear relationship, which can be seen in the correlation below. Given the high correlation between profits and market value, it is thus unsurprising that salary and market value are also not highly correlated.

```
cor(ceo.df$salary, ceo.df$mktval)
```

```
## [1] 0.4119486
```

Similar to salary and profits, there appears to be some linearity when profits and market value are low, but as market value rises the linear reltionship weakens.

```
plot(ceo.df$mktval, ceo.df$salary,
     main = "Salary vs. Market Value in 1990",
     xlab = "Market Value ($ millions)",
     ylab = "Salary ($ thousands)")
abline(lm(ceo.df$salary ~ ceo.df$mktval), col = "red")
```



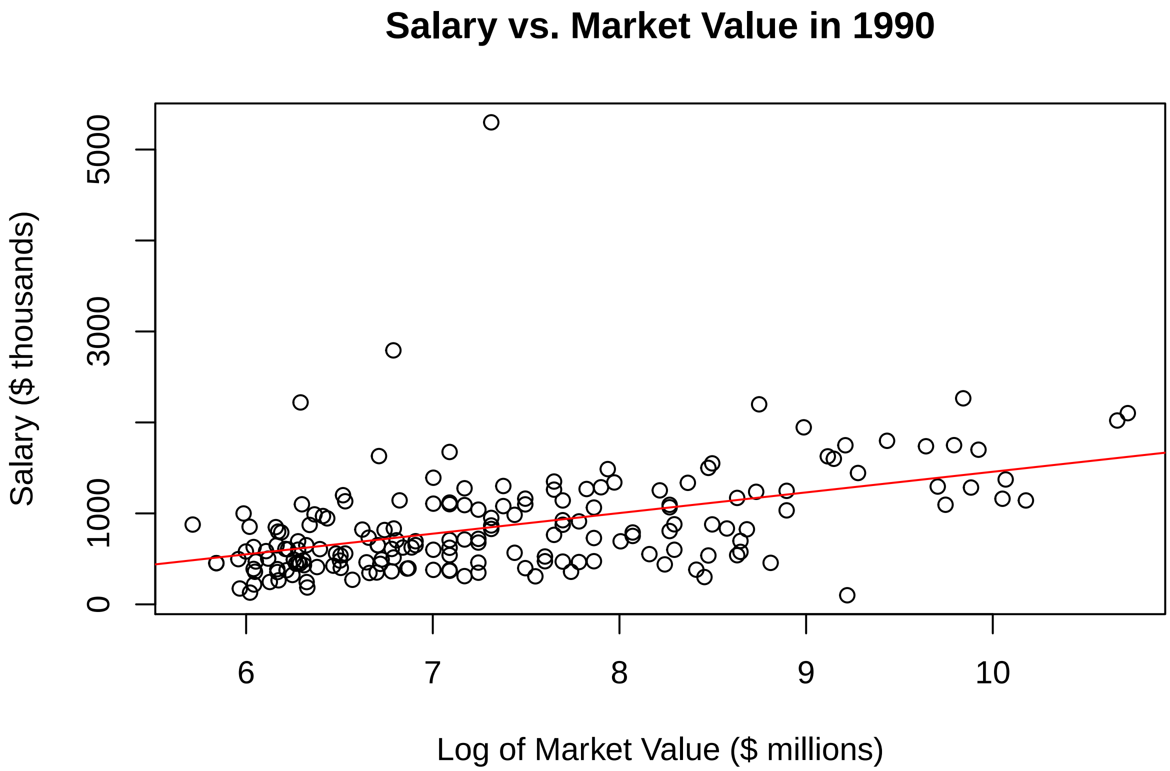
As we did with the profit variable, a logarithmic transformation on the market value variable reveals a more linear relationship between the two variables.

```
plot(log(ceo.df$mktval), ceo.df$salary,
     main = "Salary vs. Market Value in 1990",
     xlab = "Log of Market Value ($ millions)",
     ylab = "Salary ($ thousands)")
```

```
## Warning in log(ceo.df$mktval): NaNs produced
```

```
abline(lm(ceo.df$salary ~ log(ceo.df$mktval)), col = "red")
```

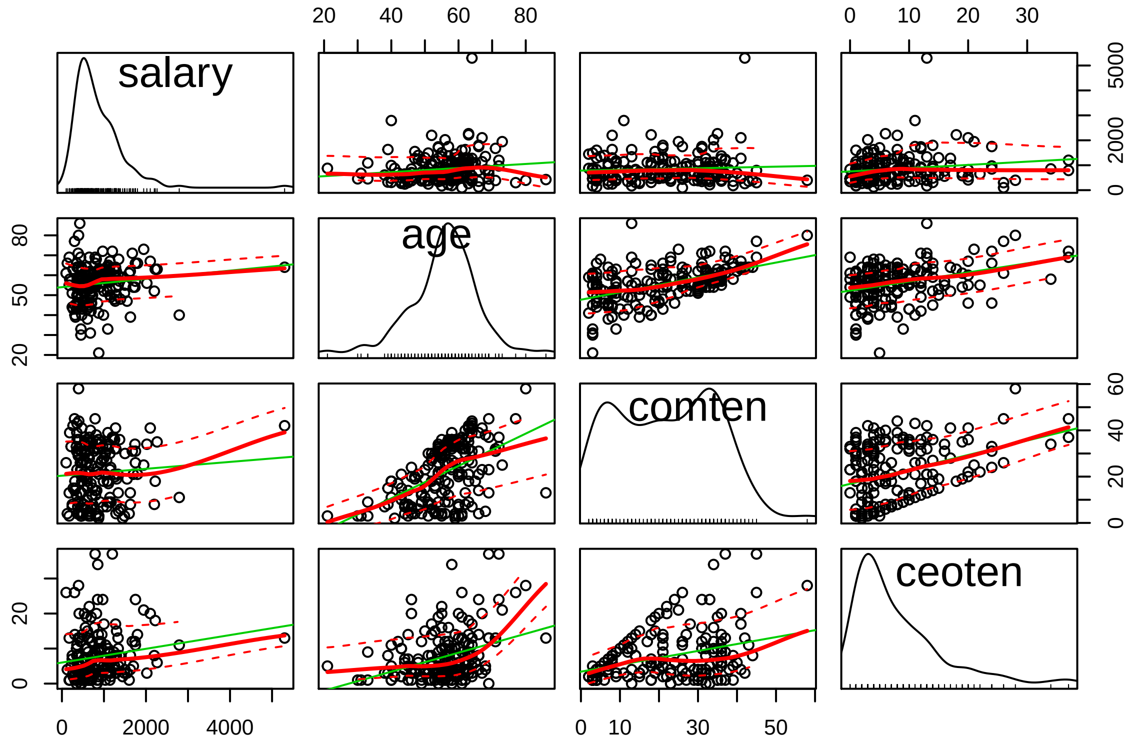
```
## Warning in log(ceo.df$mktval): NaNs produced
```



Possible Secondary Variables

Some secondary variables that may affect salary are the CEO's age, tenure, and education level. Below is a scatterplot of salary on age and tenure variables. Since education level is split into two dummy variables, we do not find a scatterplot of it to be useful.

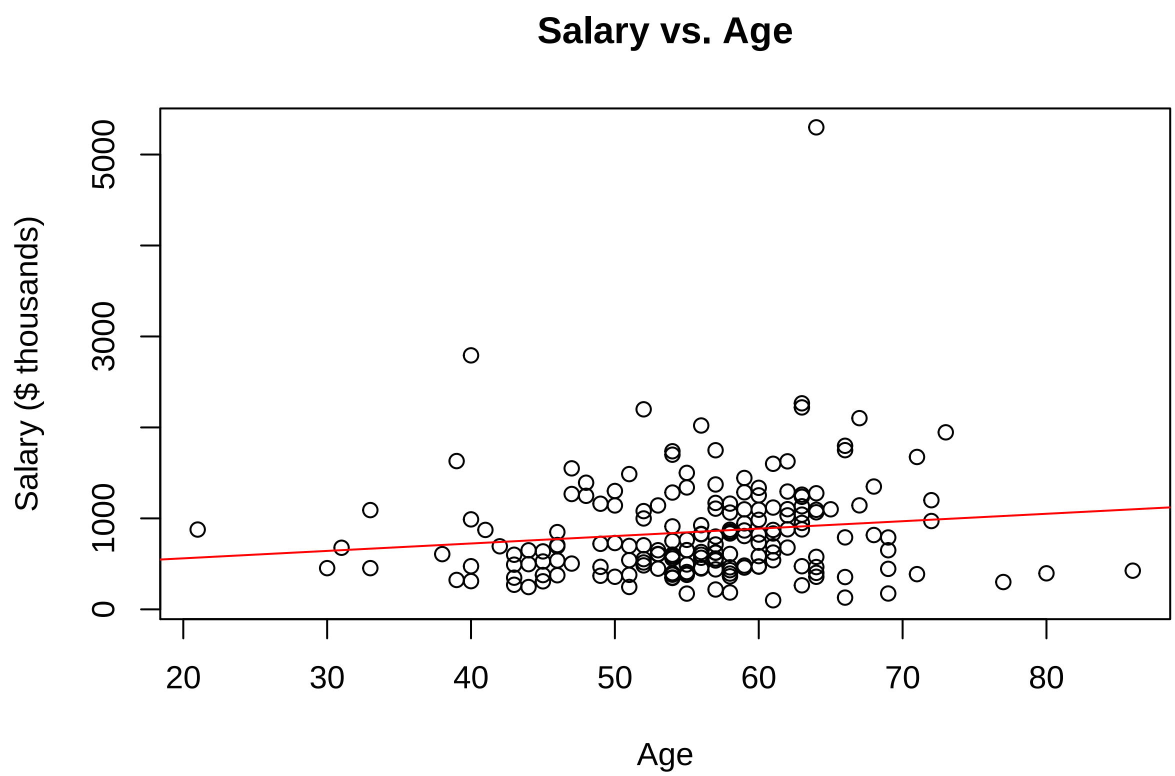
```
scatterplotMatrix(~ salary + age + comten + ceoten, data = ceo.df)
```



Salary vs. Age

Age has a slight positive relationship with CEO salary, likely because older CEOs have more work experience. However, with a correlation of only 0.130081, age is not a key variable in this analysis.

```
plot(ceo.df$age, ceo.df$salary,
     main = "Salary vs. Age",
     xlab = "Age",
     ylab = "Salary ($ thousands)")
abline(lm(ceo.df$salary ~ ceo.df$age), col = "red")
```



Salary vs. Education

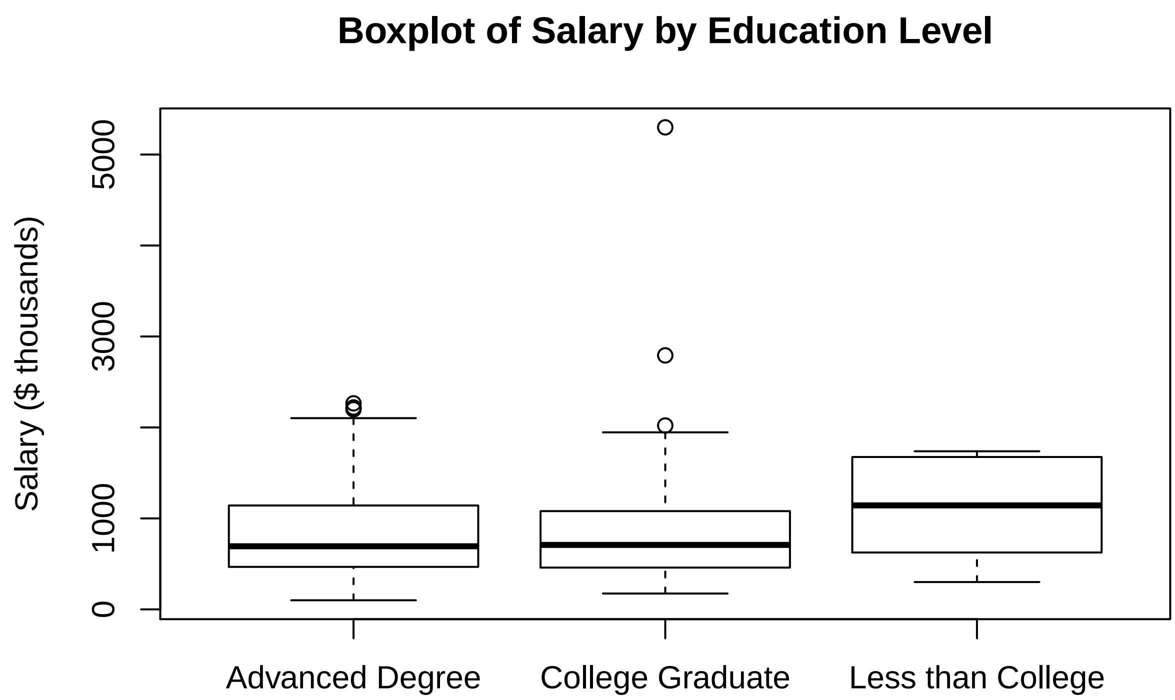
We assign each CEO to one of three education levels: Advanced Degree, College Graduate, or Less than College. There are 2 cases of CEOs with an advanced degree but no college degree, and these are assigned to the "Advanced Degree" level.

```
educLevelFunc <- function(college, grad) {

    if (grad == 1) {retStr = "Advanced Degree"}
    else if (college == 1) {retStr = "College Graduate"}
    else {retStr = "Less than College"}

    return(retStr)
}

ceo.df$educLevel <- mapply(educLevelFunc, ceo.df$college, ceo.df$grad)
table(ceo.df$educLevel)
```

```
##
##   Advanced Degree  College Graduate Less than College
##               102                78                 5
```

Following is a boxplot of CEO salary by education level. The distribution for the "less than college" category is not reliable since it only has 5 data points. However, we can see that CEOs with an advanced degree have approximately the same salary distribution as CEOs with only a college degree.

```
boxplot(salary ~ educLevel, data=ceo.df,
     main = "Boxplot of Salary by Education Level",
     ylab = "Salary ($ thousands)")
```



Salary vs. Tenure

We have two tenure variables: CEO tenure and company tenure. These variables are closely related, since company tenure must rise whenever CEO tenure rises. Note that for CEOs who are hired from outside the company, CEO tenure and company tenure are the same.

```
outside.ceo <- subset(ceo.df, ceo.df$ceoten == ceo.df$comten)
pct.outside.ceo <- (length(outside.ceo$ceoten) / length(ceo.df$ceoten))
```

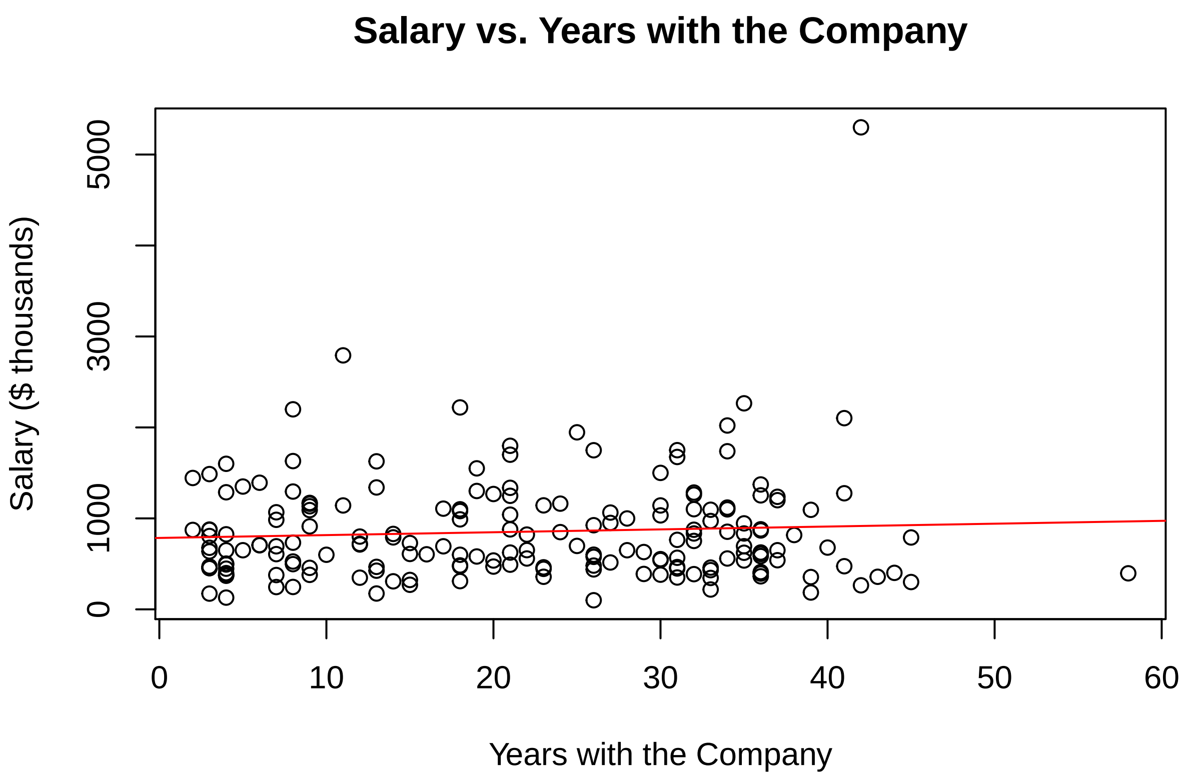
(*Note:* In this dataset, 20.5% of CEOs are brought in from outside the company.)

As shown in the scatterplots below, CEO salary has essentially no relationship with the number of years the CEO has been with the company, and only a slight positive relationship with the number of years as CEO.

```
cor(ceo.df$salary, ceo.df$comten)
```

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

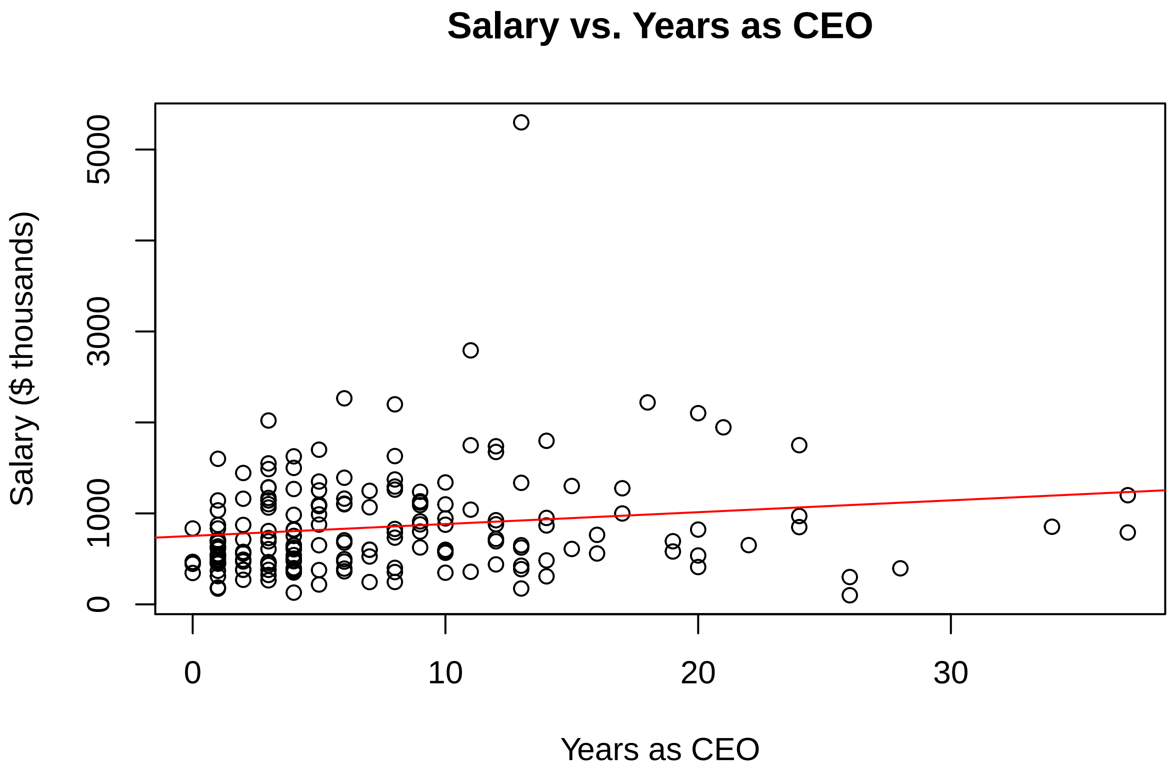
```
plot(ceo.df$comten, ceo.df$salary,
     main = "Salary vs. Years with the Company",
     xlab = "Years with the Company",
     ylab = "Salary ($ thousands)")
abline(lm(ceo.df$salary ~ ceo.df$comten), col = "red")
```




```
cor(ceo.df$salary, ceo.df$ceoten)
```

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

```
plot(ceo.df$ceoten, ceo.df$salary,
     main = "Salary vs. Years as CEO",
     xlab = "Years as CEO",
     ylab = "Salary ($ thousands)")
abline(lm(ceo.df$salary ~ ceo.df$ceoten), col = "red")
```



Potential Confounding Effects

Aaa

Variable Coding Issues and Missing Values

Mention 2 CEOs with advanced degree but not college degree
Mention 1 CEO with CEO tenure > company tenure
Mention 5 Companies with market cap equal to -1

Conclusion

Aaa