# Lab 1: Exploratory Analysis of CEO Salary Data

*Carmen Easterwood & Andrew Kabatznick*

*January 29, 2016*

## Introduction

Aaa

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

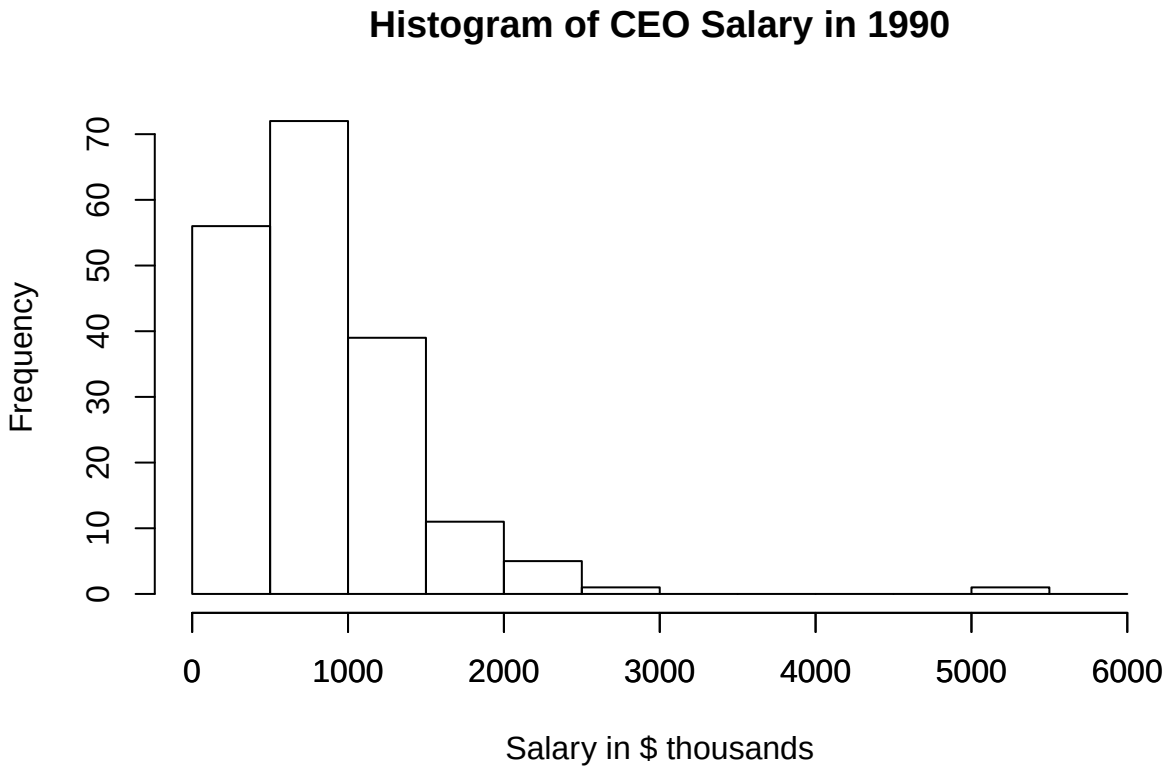
## Univariate Analysis of Key Variables

In the following section we will perform a univariate analysis of each of the variables in this dataset. Salary is our outcome variable, and profits and mktval are the key measures of market performance.

### 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 of Age is large:

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

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

### 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 are college graduates.

**Note: Should we mention there are 2 CEOs with an advanced degree but no college degree? (Can that be correct?)**

### Years With Company

Summary Statistics for Years with Company

```
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
```

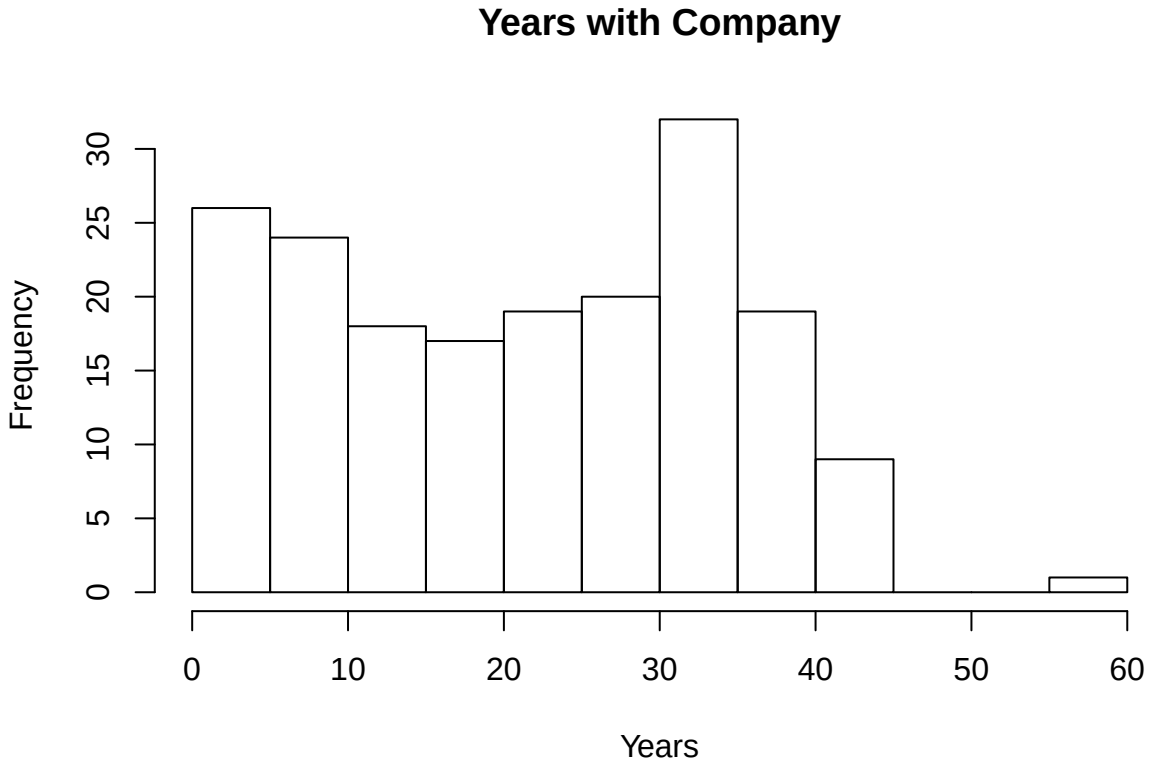
Standard Deviation of Years with Company

```
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**

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
```

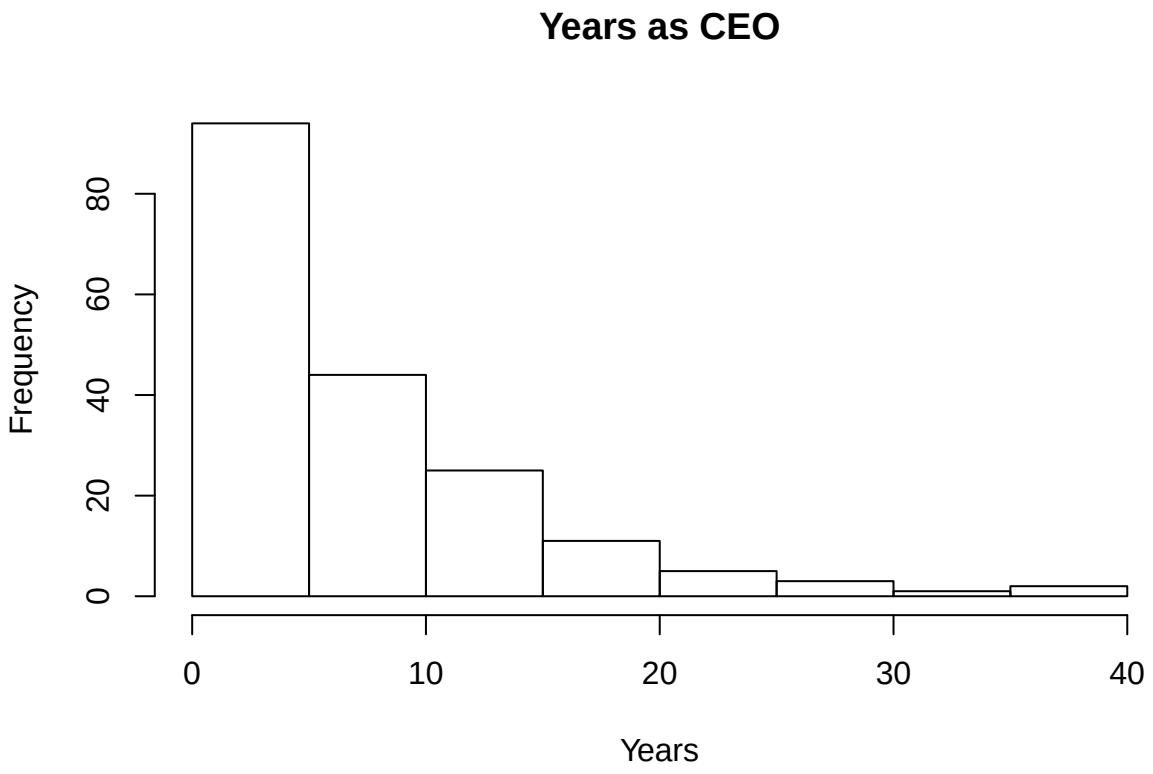
Standard Deviation of Years as CEO

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

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

Histogram of Years as CEO

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




**Profits**

Summary Statistics for Profits

```
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

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

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

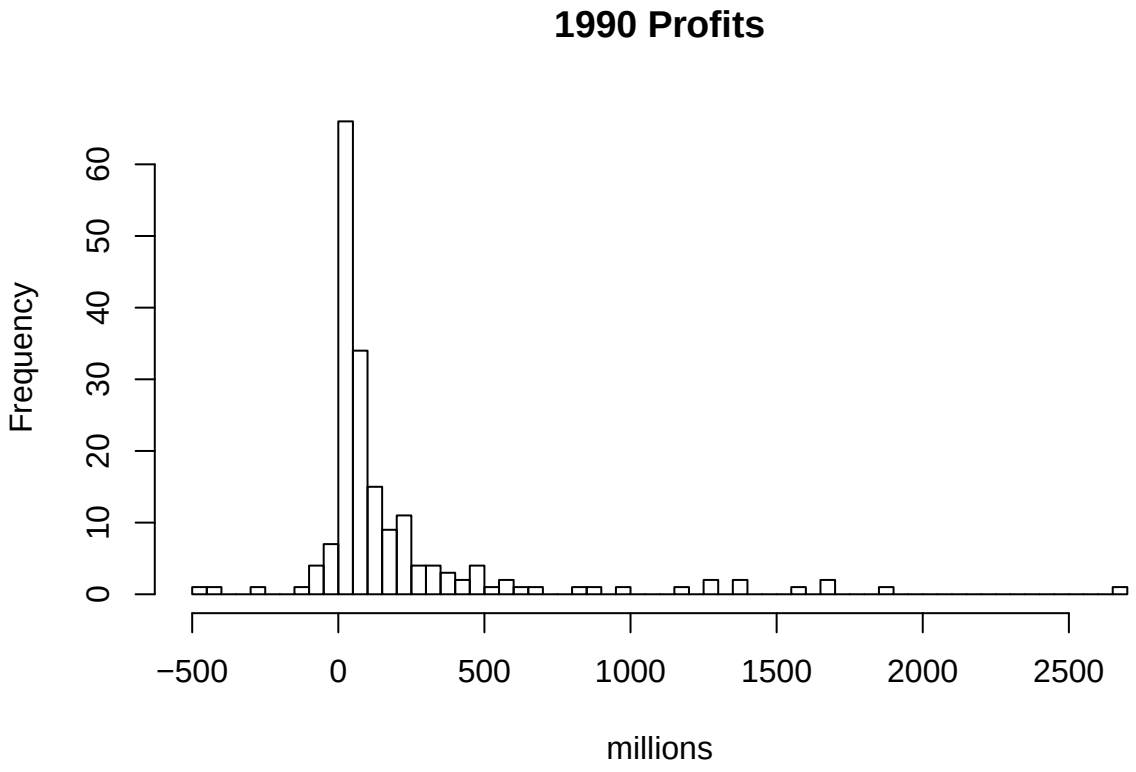
Standard Deviation of Profits

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

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

Histogram of Profits

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




**Market Value**

Summary Statistics for Market Value

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

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

Variance of Market Value

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

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

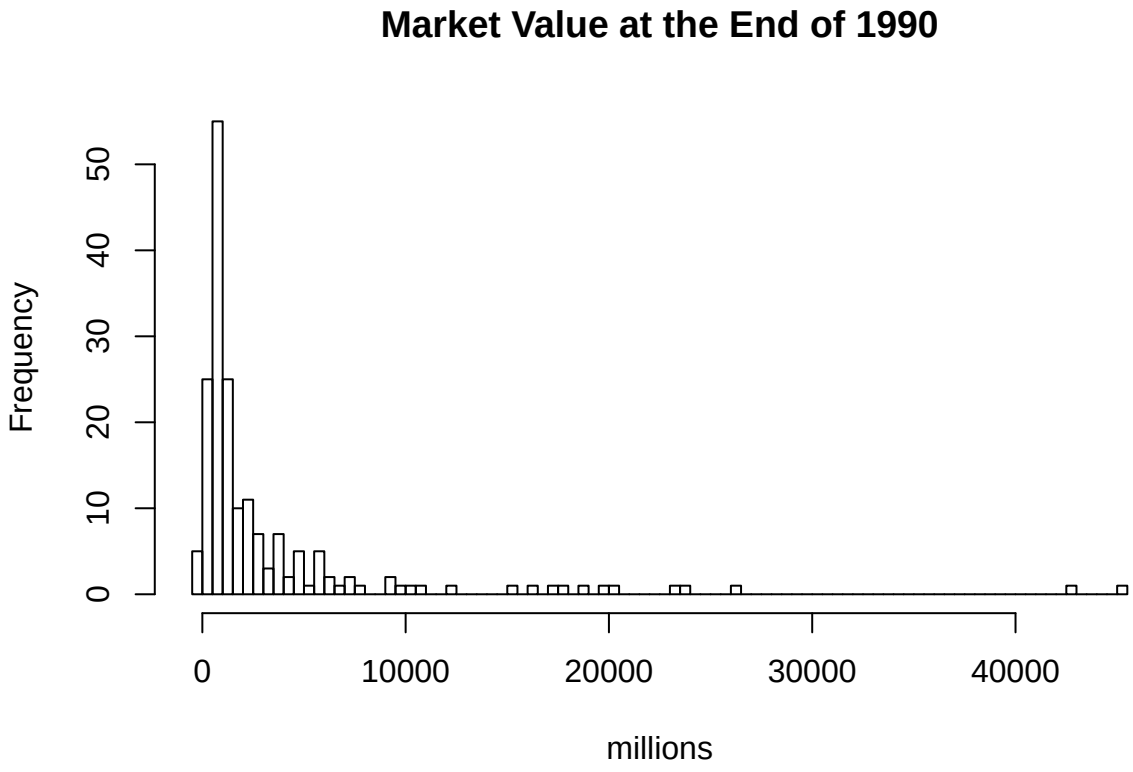
Standard Deviation of Market Value

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

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

Histogram of Market Value

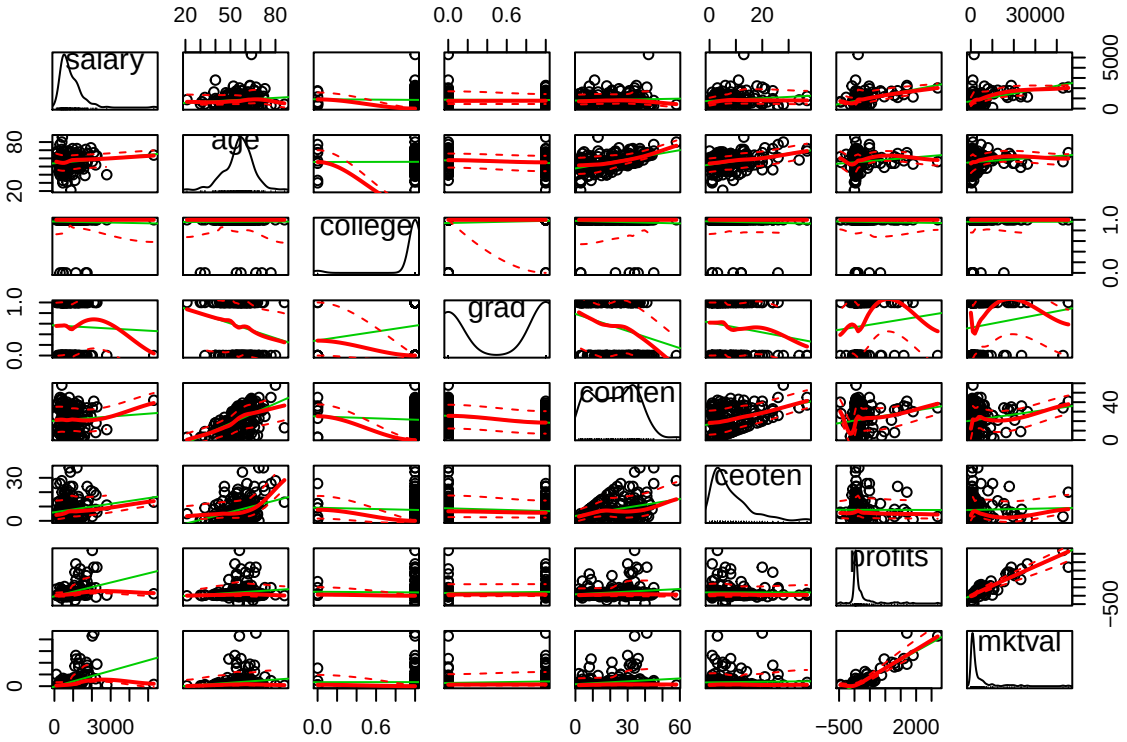
```
hist(ceo.df$mktval, main="Market Value at the End of 1990",
     xlab = "millions", breaks = 100)
```




## Key Bivariate Relationships

We will now analyze some key bivariate relationships. Below is a scatterplot matrix of the variables in this dataset.

```
library(car)
scatterplotMatrix(ceo.df)
```




### Profits & Market Value

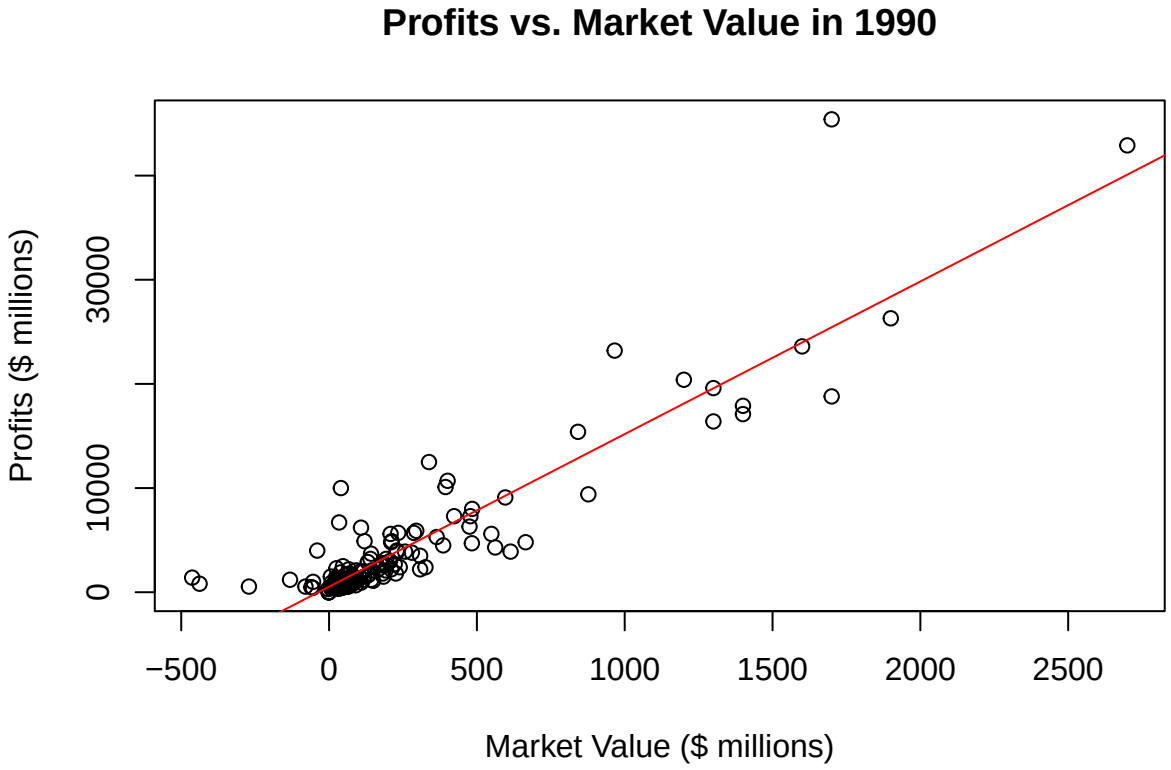
These are our two measures of company performance, so we confirm they are positively correlated.

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

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

For the most part, as profits increase, so does market value.

```
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")
```




## Possible Secondary Variables

Aaa

## Potential Confounding Effects

Aaa

## Variable Coding Issues and Missing Values

Aaa

## Conclusion

Aaa