Lab 1: Exploratory Analysis of CEO Salary Data

Anamika Sinha, Asha Anju, Noah Randolph

June 3, 2017

Introduction

An exploratory analysis of CEO and company data was conducted to explore CEO salary, with a particular interest in answering the question of whether company performance is related to CEO salary.

Exploratory data analysis techniques were used and will be presented with commentary on the findings. We first explore data in univariate fashion, followed by multivariate analyses selected for there contribution to the exploration of CEO salary. In general, CEO salary is treated as a response variable, while other datasets are treated as predictor variables.

Setup

We will utilize the "car: Companion to Applied Regression" package in R and load it, as well as the data file of interest, here.

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

Univariate Analyses

We investigate the variables independently to first understand the types of data, their frequency distributions, and to observe any anomalies, apparent data errors, and missing values. The dataset describing college attendance will not be analyzed due to the small number of CEOs who did not attend college (<4%), as can be seen below.

```
CEO$college <- factor(CEO$college, labels = c("Did Not Attend College",
                                              "Attended College"))
summary(CEO$college)
```

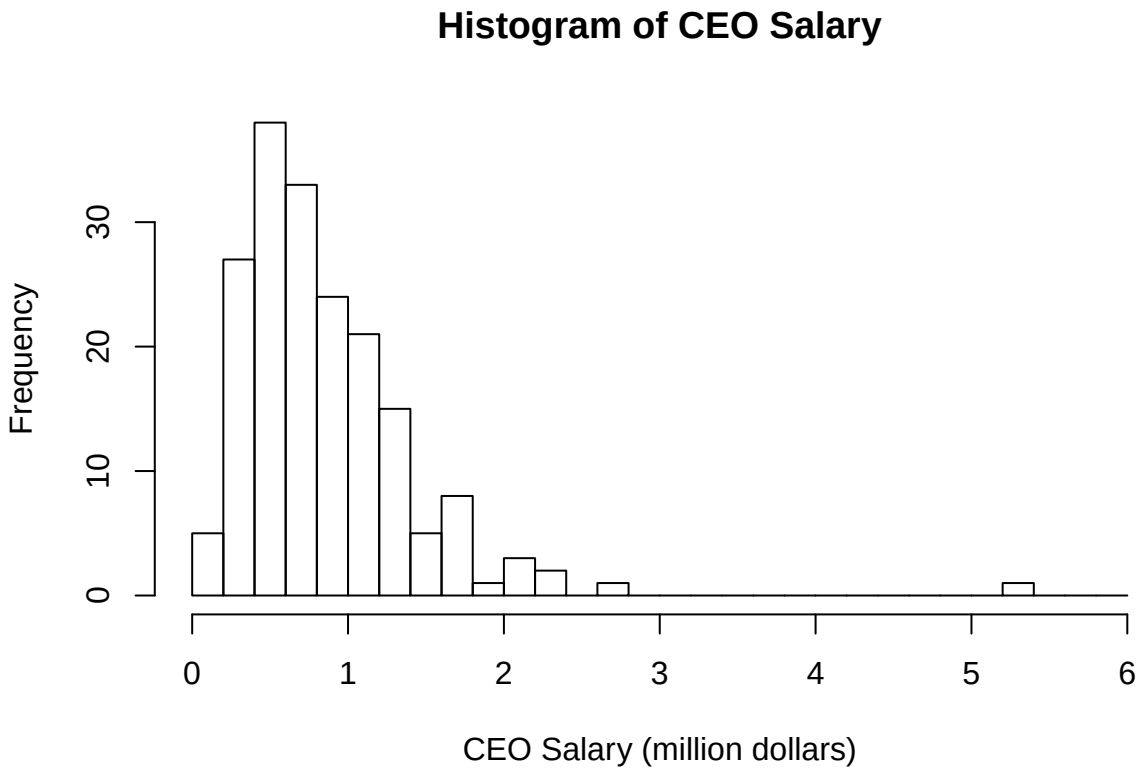
```
## Did Not Attend College       Attended College
##                      7                    177
```

Univariate Analysis of Salary Data

We start with the analysis of CEO Salary Data, our primary predictor variable of interest. We first examine the histogram of salary information.The individual salary amounts have been scaled to reflect the value in millions We have set the break points manually but the bin width is same as that obtained using Freedman and Diaconis rule.

```
CEO$salary <- CEO$salary/1000
hist(CEO$salary, breaks = seq(0,6,.2),
     xaxt = "n",
     main = "Histogram of CEO Salary",
```

```
     xlab = "CEO Salary (million dollars)")
axis(1, seq(0,6,1))
```



Next we look at the summary statistics for the salary data.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.4708  0.7005  0.8560  1.1025  5.2990
```

The following observations are apparent from our analysis of CEO salaries:

1. The distribution of CEO salaries is positively skewed with a peak around $0.6 million and a maximum value of $5.299 million.
2. Except the maximum value, all salaries are below $3.0 million. Most of the salaries are below $2.0 million.
3. The distance of maximum value from the 3rd quartile value is more than the interquartile range, thus making the data point an outlier.

Further analysis of the data corresponding to this outlier salary is required. Eventhough the maximum salary is an outlier, there is no strong evidence to conclude that it should be removed from our analysis.

Univariate Analysis of Market Value

Market value is commonly associated with company performance, which is of primary interest as it relates to CEO salary. Thus, we include an assessessment of its data in the variable, "mktval", univariately. First, a summary of the data:

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    -1.0   578.2  1200.0  3466.1  3200.0 45400.0
```

Market value data shows to be in terms of float values, including negative numbers. The range is vast, with the maximum value far greater than the values in the interquartile range. In fact, the mean is greater than not just the median, but also the 3rd quartile. Since a stock price cannot go below $0, and market value for public companies is composed of the sum of the values of each individual stock, any negative values should be scrutinized.

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

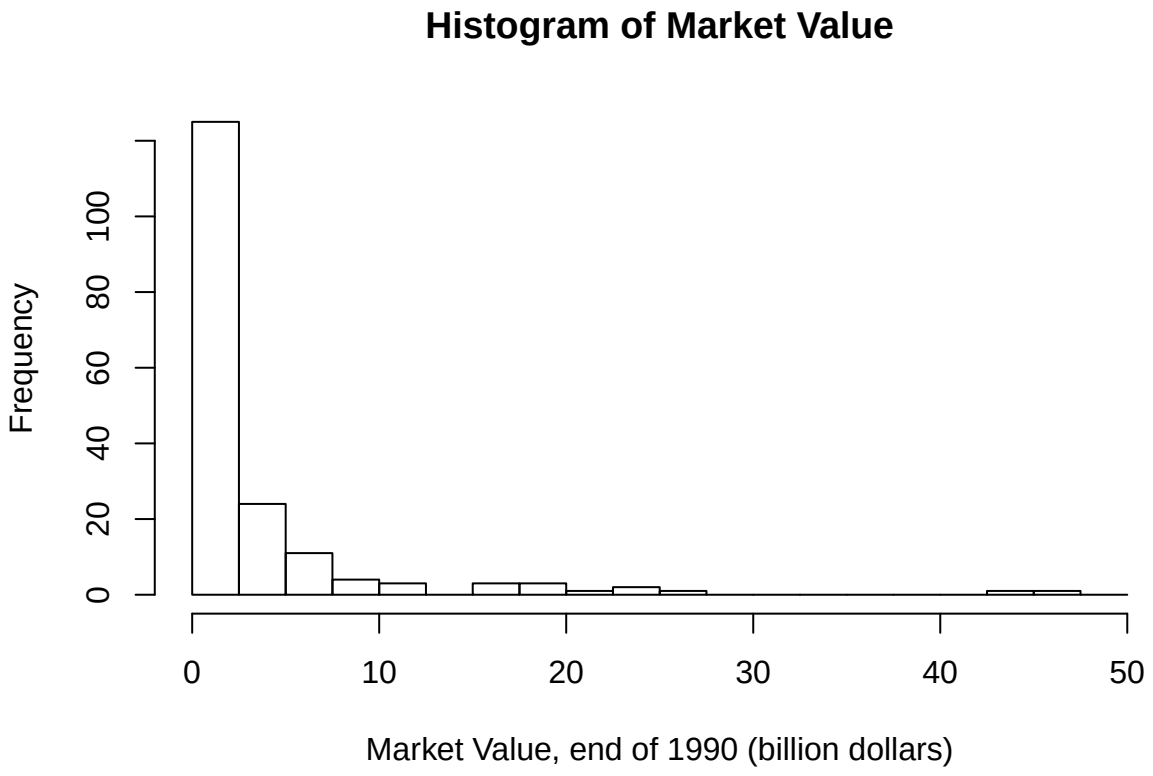
```
##     salary age          college grad comten ceoten profits mktval
## 182  0.637  45 Attended College    1      3      1      -1     -1
## 179  0.677  31 Attended College    1      3      1      -1     -1
## 180  0.173  55 Attended College    1      3      1      -1     -1
## 178  0.379  55 Attended College    1      4      2      -1     -1
## 181  0.873  61 Attended College    1      3      1      -1     -1
```

A subset of the data less than zero reveals that a coding error is apparent. The -1 values in both market value and profits likely represent data points that were unavailable for particular companies. Further analyses in this report will be on subsets of the data excluding the values of -1 for market value and profits.

```
cleanmktval <- subset(CEO$mktval, CEO$mktval > 0)
```

Next, we create a histogram of the cleaned up market value data. The dataset is transformed by a factor of 1/1,000 to make for easier reading of the X-axis. The bins are set with even widths at a number of 20, which is an amount between two common rules, one of which is the square root of the number of data points, and the other is the Freedman and Diaconis rule.

```
mktvalbillions <- cleanmktval/1000
hist(mktvalbillions, breaks = seq(0, 50, 2.5),
     main = "Histogram of Market Value",
     xlab = "Market Value, end of 1990 (billion dollars)")
```



The histogram reveals the following characteristics:

1. Market value is skewed toward the positive and frequencies drop steeply between $250 million and $500 million, for a decaying profile of frequencies as market value increases.
2. Although market value does not take on negative values, integer effects are not of concern due to the float type of data and its high dispersion.
3. Most of the market values fall below $500 million.
4. There is a large gap between highly dispersed data between roughly $2.75 billion and $4.25 billion.

Although the market values between $4 billion and $5 billion falls far outside of the bulk of the data, there is no strong evidence to conclude that they should be removed from the analysis as outliers.

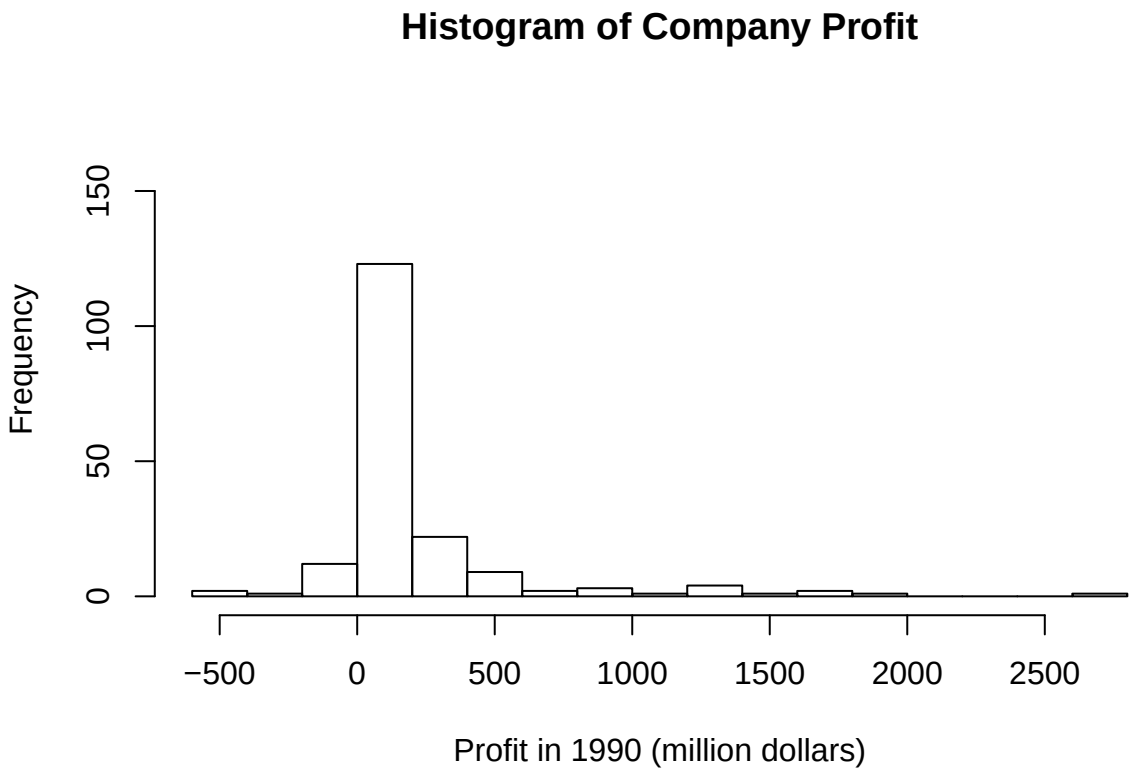
Univariate Analysis of Company profits

Next, we investigate company profits as it is also an indicator of comany performance.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -463.00   31.75   57.00  199.89  197.75 2700.00
```

```
hist(CEO$profits,
     breaks = 20,
     ylim = c(0,175),
#    labels = T,
     main = "Histogram of Company Profit",
     xlab = "Profit in 1990 (million dollars)")
```



Observations from :

1. Most companies have profits below 250 million.
2. The histogram has a slight positive skew due to the presence of some outliers.They seem to be pushing the mean to even beyond the 3rd quartile. The mean is much higher than the median value.
3. 15 companies with negative profits is worth investigating. Since most of the profits is from zero to 250 million that bin is worth investigating.
4. It will be interesting to see the salaries of CEOs with negative profit for 12 companies in 0 and -200 million range. We can ignore 5 values with profits == -1 for that analysis as the mkatval analysis suggests a coding error.

Univariate Analysis of CEO Age

We next examine the age of CEO's to understand the distribution and its features.

```
summary(CEO$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   51.00   57.00   55.98   61.25   86.00
```

```
hist(CEO$age, breaks = "FD",
     main = "Histogram of CEO Age",
     xlab = "Age of CEO (years)")
```

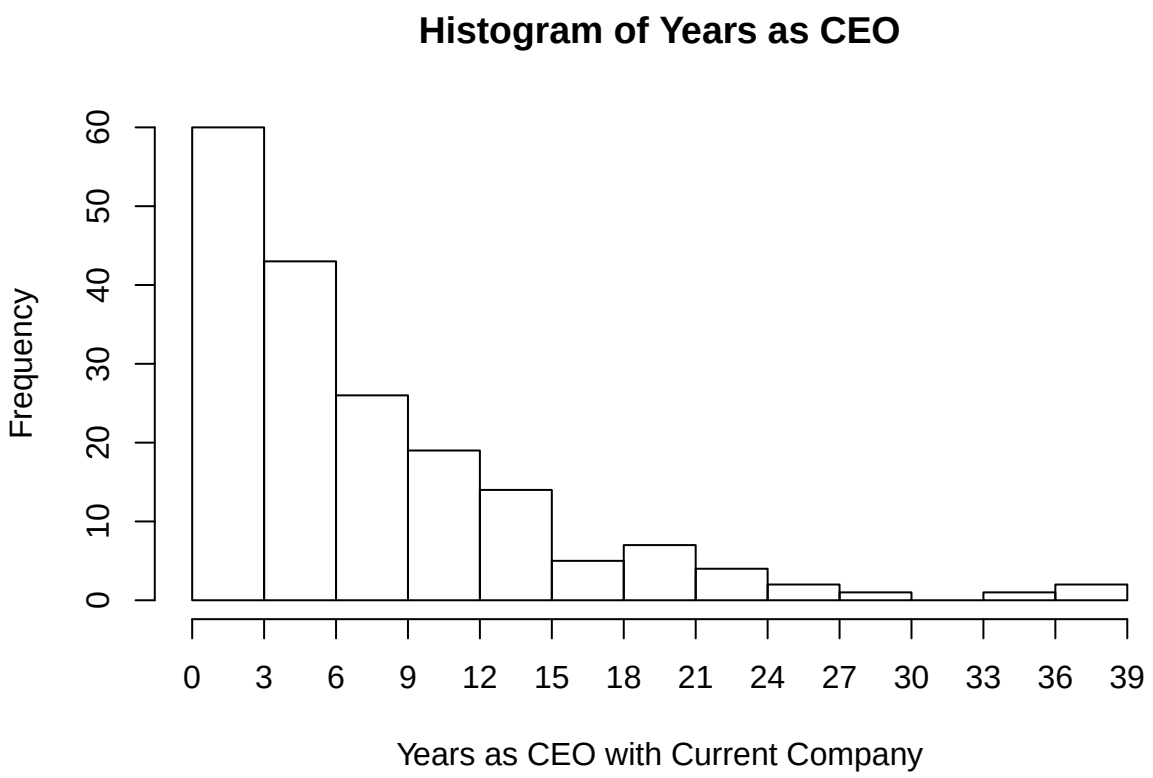
The following observations are apparent from our analysis of age information:

1. The ages cover a large range with a minimum age of 21 years and maximum age of 86 years.
2. The distribution peaks peak around 55-60 years of age and have a slight negative skew.
3. 50% of the CEO's in the dataset have ages in a 10 year range between 51 and 61.25 years.

Univariate Analysis of CEO Tenure

We next examine years of CEO tenure, which may relate to CEO salary as any normal salary typically grows with tenure. We manually set the cut points for our histogram to make 13 bins, which is roughly the square root of the number of data points.

```
hist(CEO$ceoten, breaks = seq(0, 39, by=3), main = "Histogram of Years as CEO", xlab = "Years as CEO wi
axis(1, seq(0, 39, by=3))
```



The following observations are apparent in the histogram of CEO tenures:

1. The lower bound is at zero years (indeed, negative years would not make sense in this context).
2. The data is skewed positively, with more than one data point in the rightmost bin of 36 to 39 years.
3. CEO tenure frequencies peak in the lowest number of years and then mostly drop in a decaying form as the number of years increases.
4. There is a gap in tenures from 30 to 33 years and a decrease from 15 to 18 years, but these are most likely due to dispersion of lower frequencies of higher year values.

Univariate analysis of Company tenure

We analyzed the total number of years CEOs stayed with the company including years before or after becoming CEO.

```
summary(CEO$comten)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   21.50   21.77   33.00   58.00
```

```
hist(CEO$comten,
#     labels = T,
     ylim = c(0,40),
     main = "Histogram of Years in Company",
     xlab = "Years"
     )
```

Observations:

1.We see that there is almost a uniform spread of data across the number of years from zero to forty years.

2.There is a bit of spike at 30-35 years and 1 outlier in the 55 to 60 years.

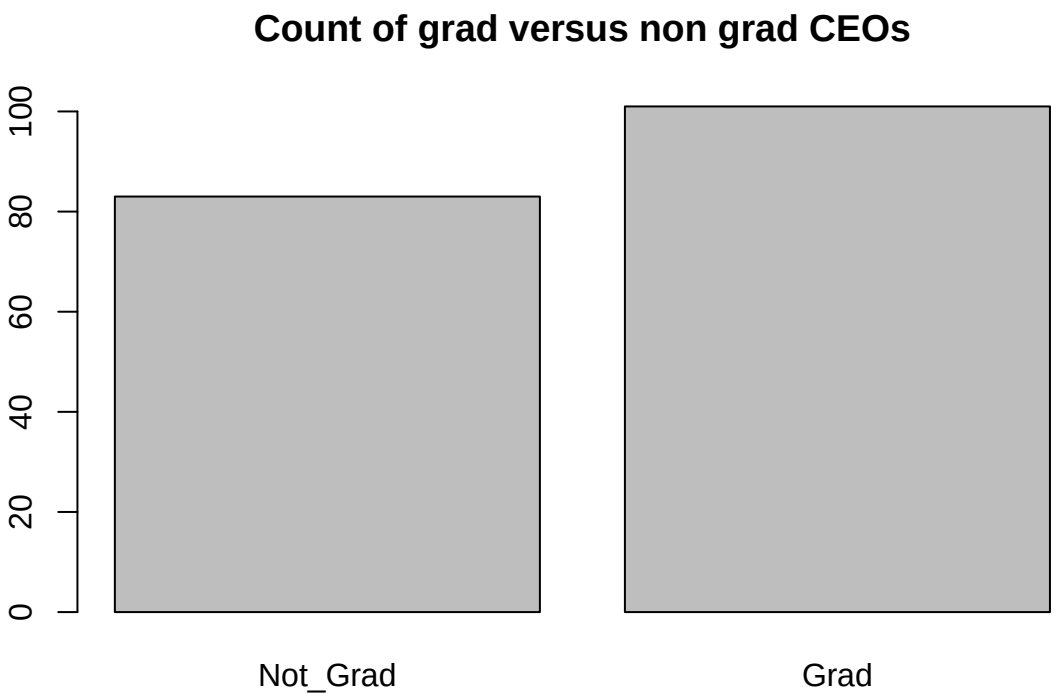
3.Overall, company tenure does not seem to have an impact on the output.

Univariate analysis of grad (CEOs with graduate degrees)

Since grad is a categorical variable, we create a factor variable and assign levels of GRAD and "Non_Grad" to it.

```
CEO$grad <- factor(CEO$grad,
                   levels = c("0", "1"),
                   labels = c("Not_Grad", "Grad"))
#CEO$grad
```

```
plot(CEO$grad,
     main = "Count of grad versus non grad CEOs")
```




```
summary(CEO$grad)
```

```
## Not_Grad     Grad
##       83      101
```

Observations:

1. We see that the number of CEOs without a graduate degree is pretty comparable to the number of CEOs with graduate degree.
2. This means that the data is pretty well distributed and graduate degree should not be having a huge impact on the outcome.

Multivariate Analyses

Now that we understand the type and general nature of the data in each variable, we turn to multivariate analyses to begin to answer the question of whether CEO salary is related to company performace.

Multivariate Analysis of Market Value Versus CEO Salary

Market value is plotted against CEO salary as it is one measure of company performance. We are interested in seeing the degree to which the level of CEO salary can at all be related to the market value of that CEO's company.

We don't want to include the -1 values found in the univariate analysis of market values. Conveniently, we plot the $\log_{10}$ of market value, so the -1 values and their associated salary data do not appear on the plot, though R will give a warning that NaNs are produced.

```
plot(log10(CEO$mktval), CEO$salary, main = "Market Value Versus CEO Salary",
     xlab = expression(paste(Log[10],
```

```
                              " of Market Value, end of 1990 (million dollars)")),
    ylab = "Salary (million dollars)")
```

```
## Warning in plot(log10(CEO$mktval), CEO$salary, main = "Market Value Versus
## CEO Salary", : NaNs produced
```