Lab 1

Saul Grimaldo, Melwin Poovakottu, Anirudh Mittal

1/18/2017

```
library(ggplot2)
# Assumes that we are currently in the same directory as the data set.
load("ceo_w203.RData")
```

Introduction

We were tasked to determine whether company performance is related to a CEO's salary. To answer this question, we will explore a CEO data set containing the salaries of 185 CEOs.

We operationalize the concept of performance through the use of profits as a proxy for performance.

We will examine whether we can say that a higher salary is related to better company performance. We will also explore how different variables might potentially interact with both salary and company performance to better understand the true nature of the relationship between salary and company performance.

Data Modification, Data Encoding, and Dubious Data Points

For this analysis, we assumed that if a CEO reported completing a graduate degree, they must have also completed college. This affected 2 data points in the CEO data set.

```
# counts the number of CEOs who reported having a graduate level education but no college education
nrow(subset(CEO, grad == 1 & college ==0))
```

```
## [1] 2
```

```
# updates college education for CEOs who reported having a graduate level degree.
CEO$unaltered_college = CEO$college
CEO$college = ifelse(CEO$grad == 1, 1,CEO$unaltered_college)
```

Furthermore, we encountered cases where market value and profits are set to -1. In the case of profits, this is acceptable, but only 5 data points showed a market value of less than 0 and they were all -1. Furthermore, we see that profits are -1 only when market value is also -1. As such, we updated these values to be NA.

```
nrow(subset(CEO, mktval == -1))
```

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

```
nrow(subset(CEO, profits == -1))
```

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

```
nrow(subset(CEO, mktval == -1 & profits == -1))
```

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

```
CEO$unaltered_mktval = CEO$mktval
CEO$unaltered_profits = CEO$profits
CEO$mktval = ifelse(CEO$unaltered_mktval == -1, NA, CEO$unaltered_mktval)
CEO$profits = ifelse(CEO$unaltered_mktval == -1 & CEO$profits == -1, NA, CEO$profits)
```

Finally, we also noticed that 1 observation exists where ceoten > comten. The data point also belonged to someone who reported being 21 years old, completing both an undergraduate and graduate degree, and being a CEO since he/she was 16. This data point is dubious, so we eliminated it from the data set.

```
print(subset(CEO, ceoten > comten))
```

```
##     salary age college grad comten ceoten profits mktval unaltered_college
## 183    877  21       1    1      3      5      -3    303                 1
##     unaltered_mktval unaltered_profits
## 183              303                -3
```

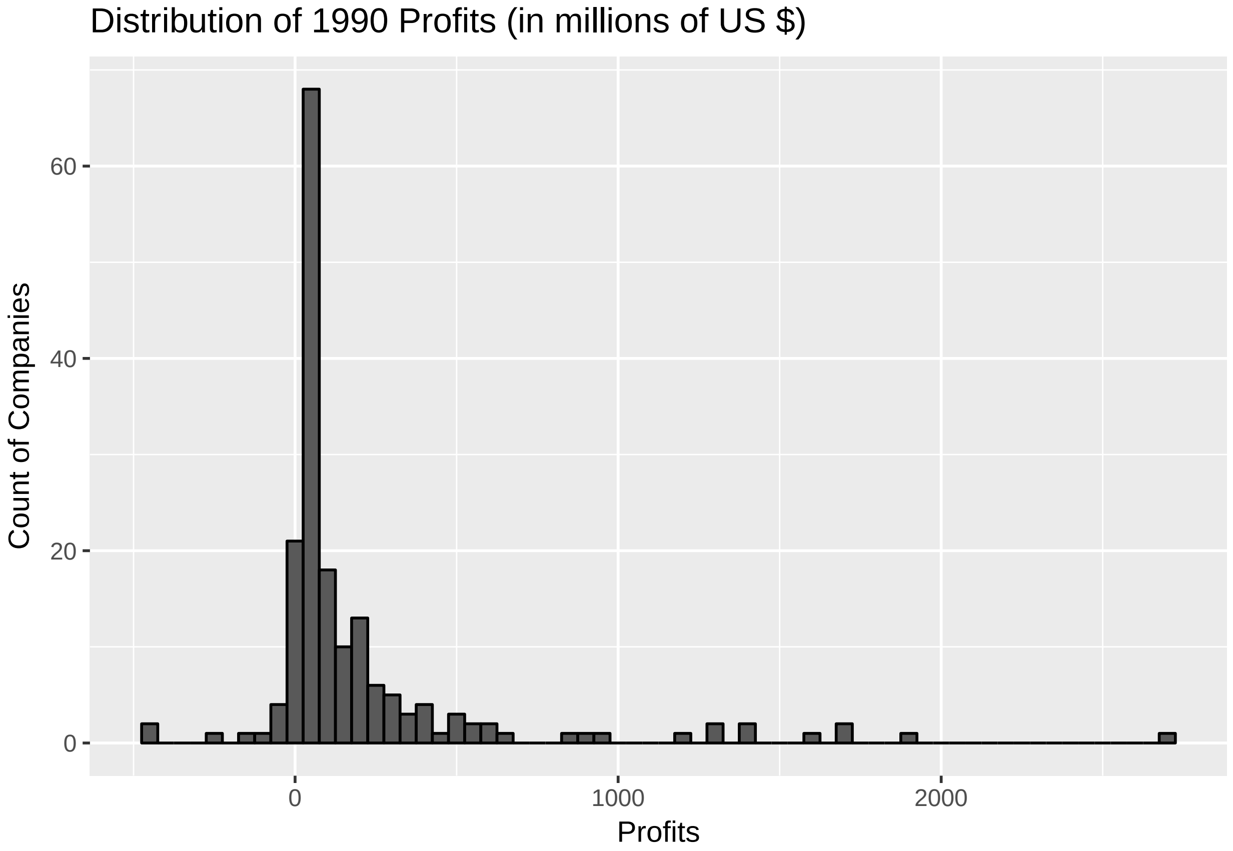
```
CEO = subset(CEO, comten >= ceoten)
```

Univariate Data Analysis

Exploration of profits

We begin our analysis by looking at profits, the variable of interest.

```
ggplot(CEO,aes(x = profits)) +
  geom_histogram(color=as.factor(1), binwidth = 50, na.rm = T) +
    ggtitle("Distribution of 1990 Profits (in millions of US $)") +
     labs(x = "Profits", y = "Count of Companies")
```



Profits, are highly skewed to the right with the majority of CEO's working at companies making between $50 million and $100 million in profits in 1990.

We thus see that most of the companies we are studying are largely fairly successful companies.

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

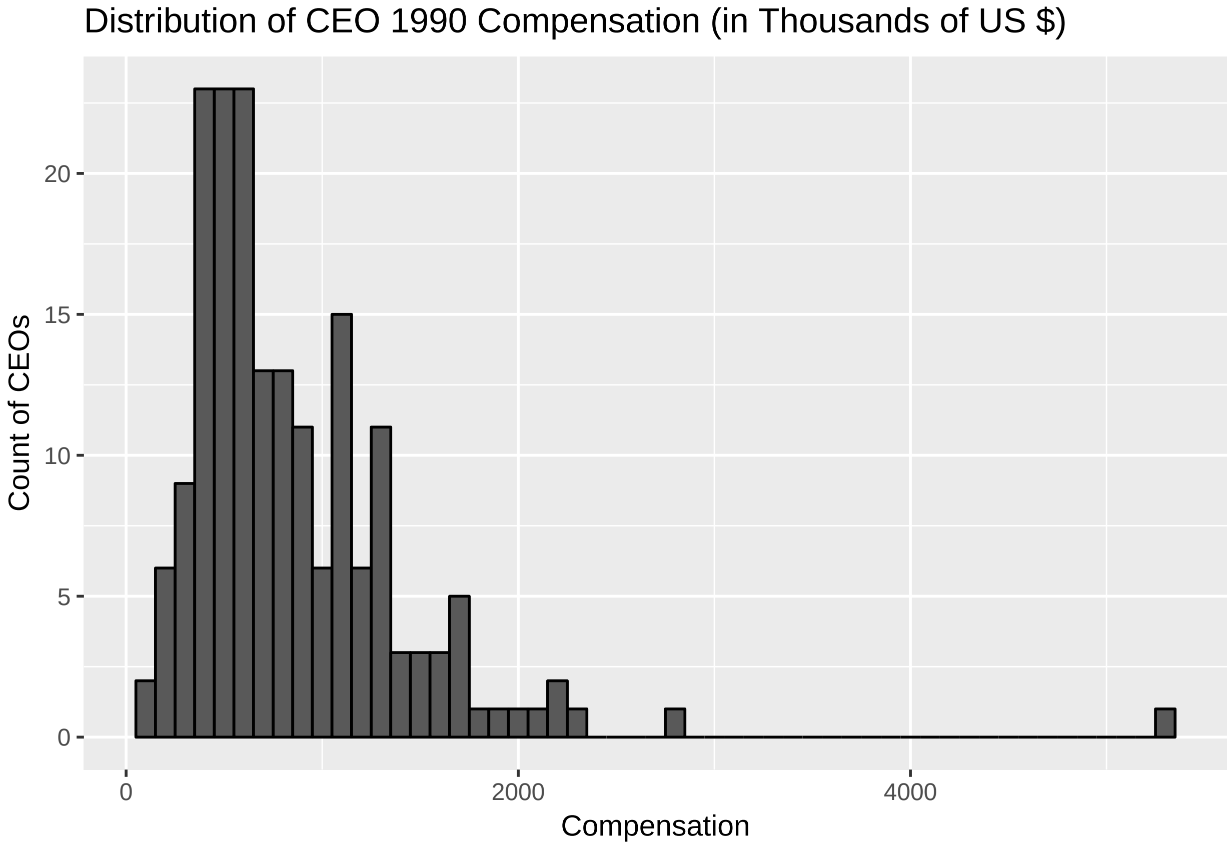
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  -463.0    34.0    63.0   205.9   207.0  2700.0       5
```

We see a fascinating phenomenon here. Due to the large skew in the data, the mean of profits is closer to the third quartile of profits than it is to the median.

Exploration of Salary

We next explore salary.

```
ggplot(CEO,aes(x = salary)) +
  geom_histogram(color=as.factor(1), binwidth = 100) +
   ggtitle("Distribution of CEO 1990 Compensation (in Thousands of US $)") +
    labs(x = "Compensation", y = "Count of CEOs")
```



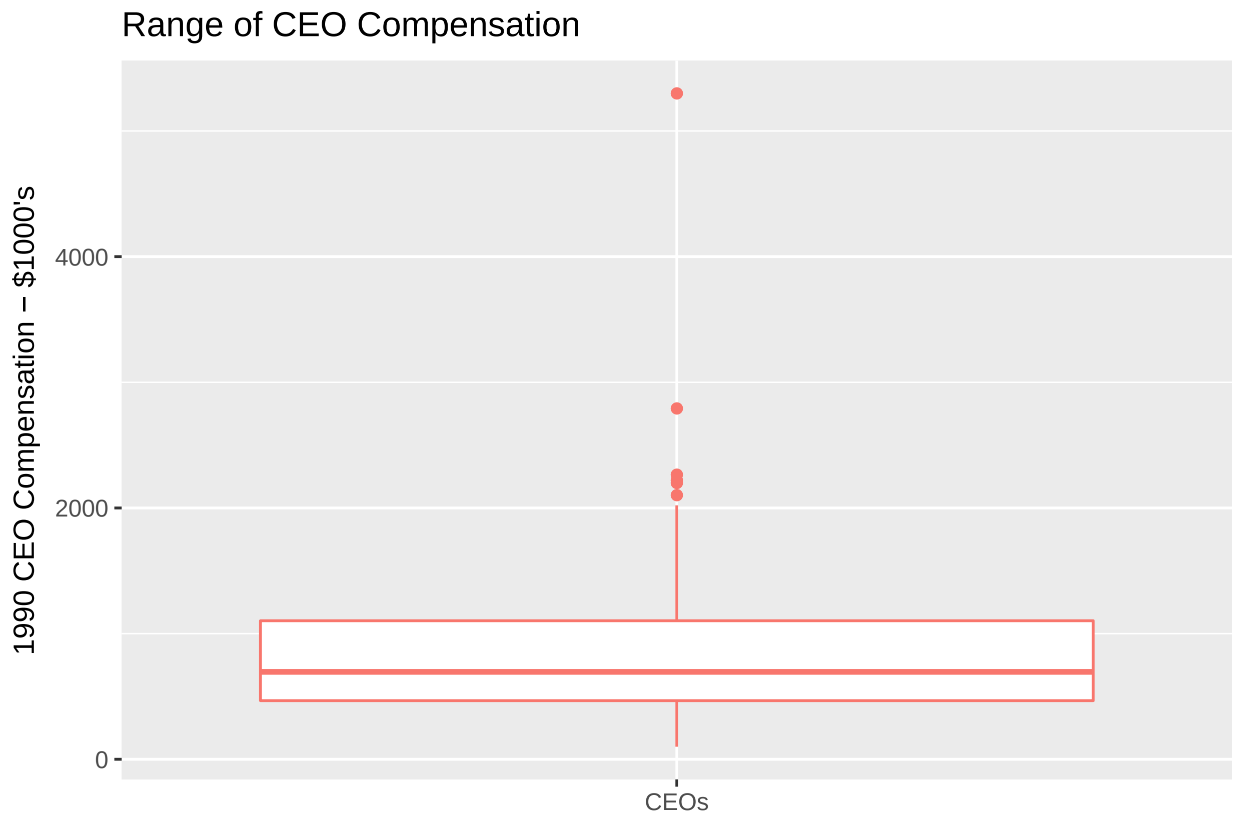
We can see that salary, like profits, has a strong right skew. However, data is more dispersed.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   466.5   695.5   852.7  1102.0  5299.0
```

We can also see that the salary varies fairly dramatically with the first quartile being at $467,000 and the third quartile being at $1,101,00.

```
ggplot(CEO, aes(x = "CEOs", y = salary, color = as.factor(1))) +
  geom_boxplot() + ggtitle("Range of CEO Compensation") +
  labs(x = "", y = "1990 CEO Compensation - $1000's" ) + theme(legend.position="none")
```



There's also a fair amount of outlier CEOs earning more than $2 million.

Exploration of Education

Another interesting variable to explore is the education of the CEO.

The college variable is a binary variable. This means that if college is 1, then the CEO completed college. Otherwise college is set to 0. Within this sample, we see that more than 97% of CEO's have a college education

```r
# Calculates percent of CEOs in our data set with a college education
number_of_college_CEO = sum(CEO$college)
data_set_size = nrow(CEO)
number_of_college_CEO / data_set_size
```

```
## [1] 0.9728261
```

Only 5 CEOs in our data set did not report having a college education.

```r
data_set_size - number_of_college_CEO
```

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

Next, we look at graduate school completion. Within our sample, we see that a total of 55 % of CEO's reported having a graduate level education

```
sum(CEO$grad) / nrow(CEO)
```

```
## [1] 0.548913
```

Because of the minimal amount of data in our data set for CEOs with no education, we will focus education related questions on graduate level education.

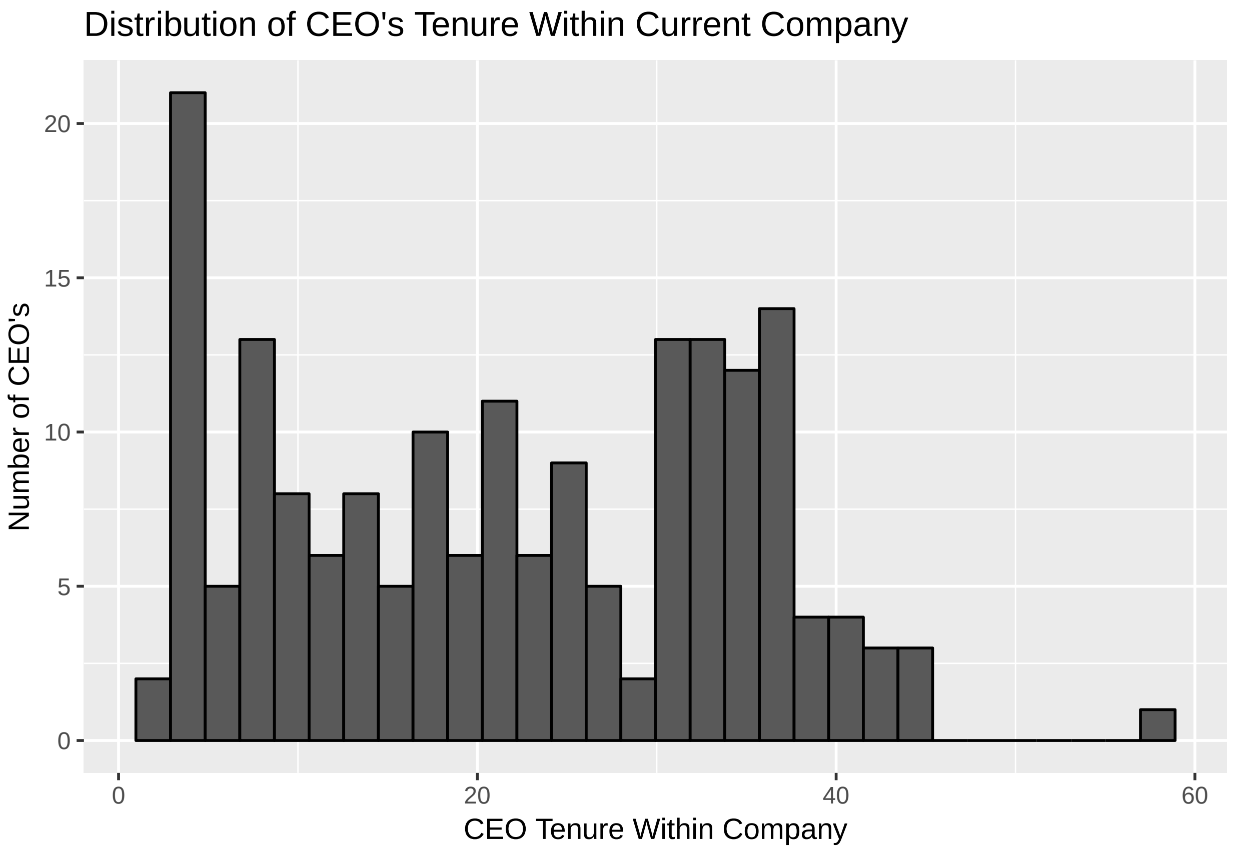
```
# Turns grad into a yes vs no variable to make it more interpretable in future analyis.

CEO$unaltered_grad = CEO$grad
CEO$grad = as.factor(ifelse(CEO$unaltered_grad == 1, "Yes", "No"))
```

Exploration of company tenure

Next, we'll explore the company tenure variable.

```
ggplot(CEO, aes(x = comten)) + geom_histogram(color = as.factor(1), bins = 30) +
  ggtitle("Distribution of CEO's Tenure Within Current Company") +
  labs(x = "CEO Tenure Within Company", y = "Number of CEO's")
```



We can see a large spike in the number of CEOs with low amounts of company tenure. This likely corresponds to CEOs who were hired from out of the company.

There is also a large bump in the number of CEOs who had 30 - 36 years of company tenure. Perhaps this is a signal that companies value promoting from within as employees who know the business well, which is something an employee will accomplish with experience working at the same business for a long period of time, will be able to make the best decisions.

Finally, we can see the number of CEOs with more than 36 years of experience dropping off dramatically. This is likely driven by CEO's entering retirement.

```
subset(CEO, comten == 58)
```

```
##     salary age college grad comten ceoten profits mktval unaltered_college
## 122    396  80       1   No     58     28      53    963                 1
##     unaltered_mktval unaltered_profits unaltered_grad
## 122              963                53              0
```

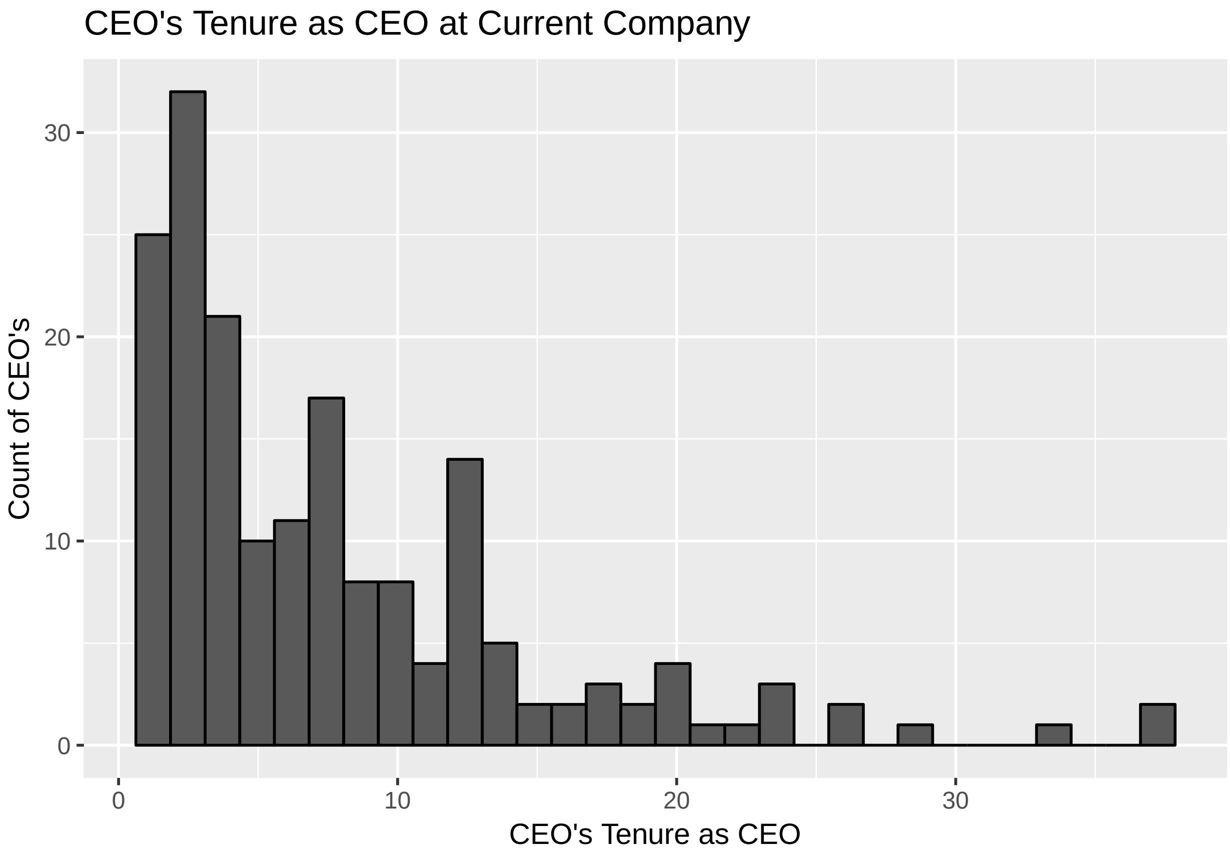
Below, we can see the range in CEO company tenure going from 2 to 58, with the mean and median being relatively close at 21.66 and 21 respectively.

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

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

Exploration of Tenure of CEO as a CEO

```
ggplot(subset(CEO,ceoten > 0),aes(x = ceoten)) +
  geom_histogram(color=as.factor(1), bins = 30) +
  ggtitle("CEO's Tenure as CEO at Current Company") +
  labs(x = "CEO's Tenure as CEO", y = "Count of CEO's")
```



The CEO's tenure has a right skew with 75% of CEO's having a tenure of less than 11 years.

As expected from a righward skew distribution the median and mean are fairly different as well.

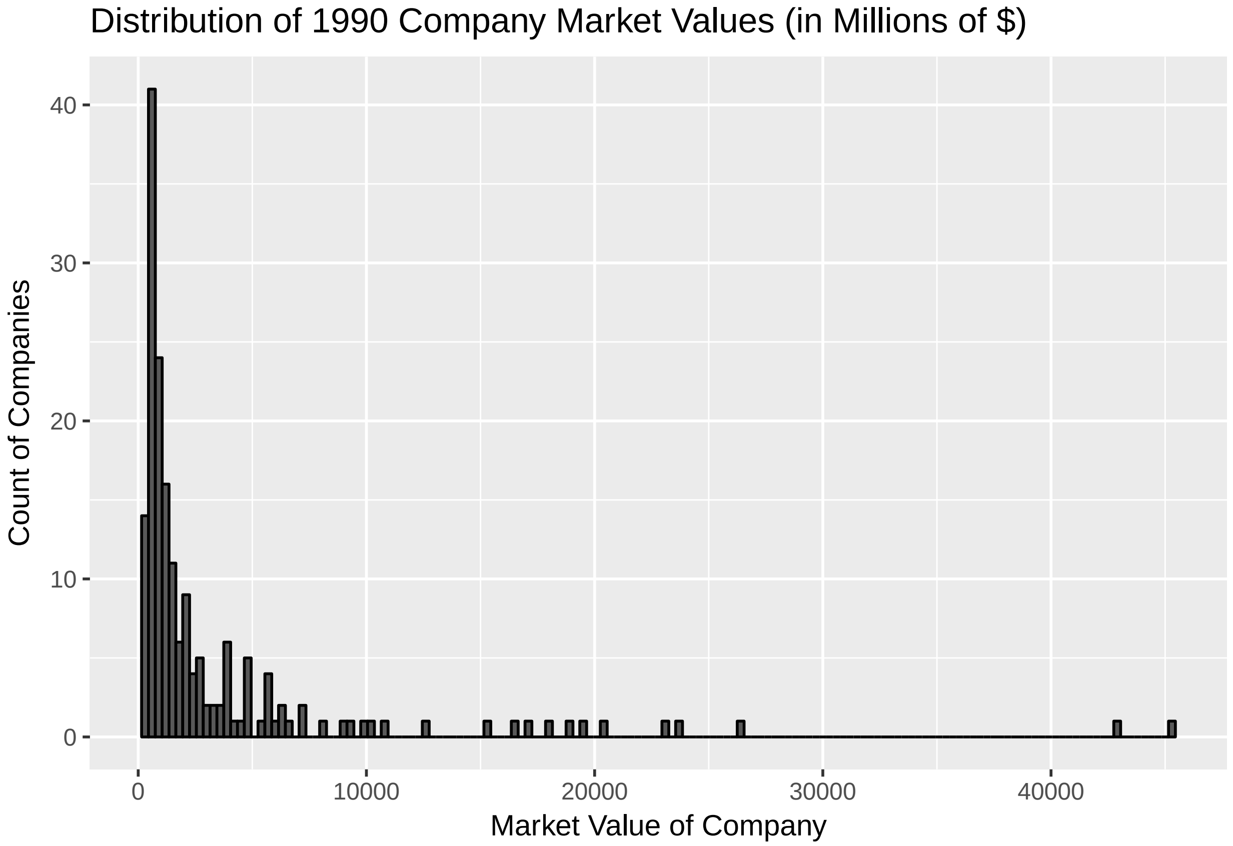
```
summary(CEO$ceoten)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   3.000   5.000   7.696  11.000  37.000
```

Market Value Exploration

We now explore market value.

```
ggplot(CEO,aes(x = mktval))+ geom_histogram(binwidth = 300, color = as.factor(1), na.rm = T) +
  ggtitle("Distribution of 1990 Company Market Values (in Millions of $)") +
   labs(x = "Market Value of Company", y = "Count of Companies")
```



Like salary and profits, we see a strong rightward skew in our data for market value.

In this case, market value, similar to profits, is highly clustered. However it is clustered in the $1 billion - $2 billion market value range.

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

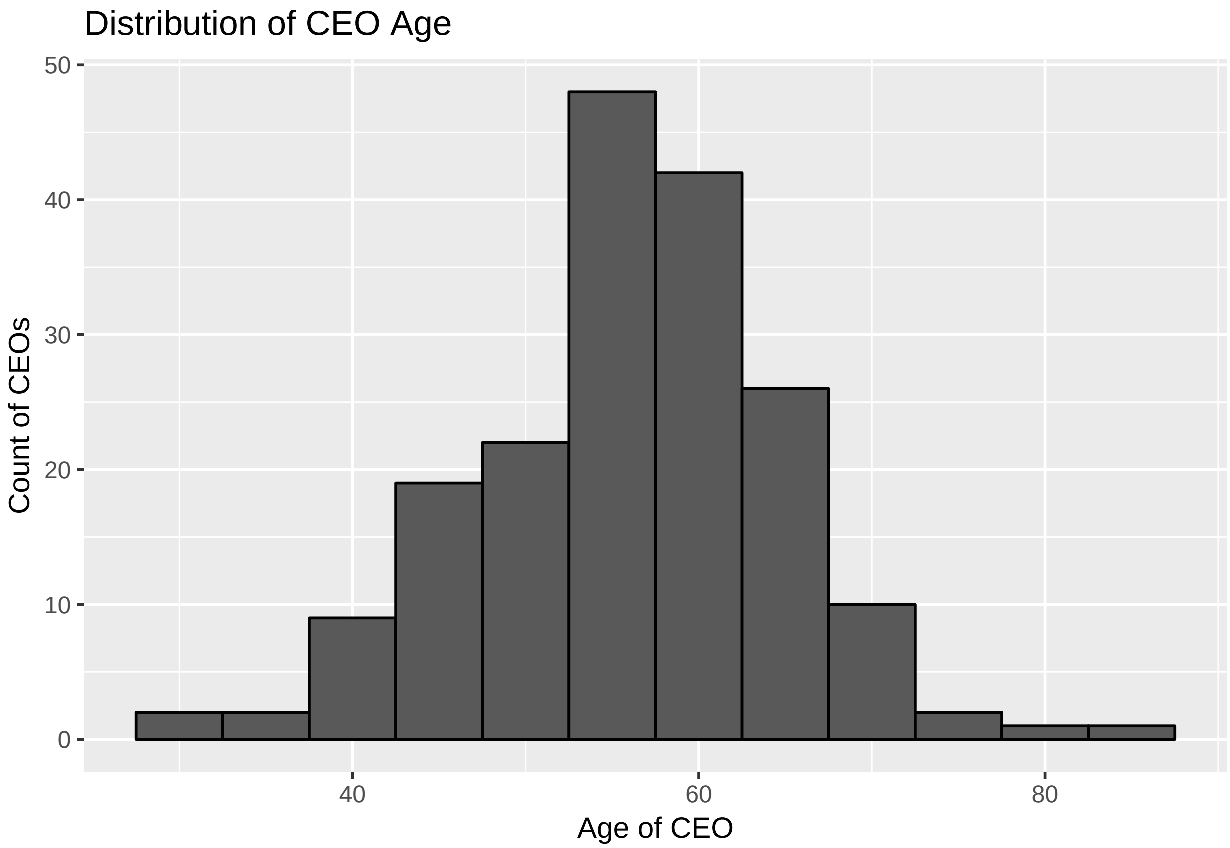
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     344     616    1200    3564    3350   45400       5
```

As expected, due to the large skew in the market value, we see that the mean and median are not close close and the interquartile range is over $2,900,000.

Exploration of Age

Looking at the histogram of age, we can see that the variable is close to a normally distributed curve with a drop after the age of 65, the age of retirement in the US.

```r
ggplot(CEO,aes(x = age))+ geom_histogram(binwidth = 5, color = as.factor(1)) +
  ggtitle("Distribution of CEO Age") + labs(x = "Age of CEO", y = "Count of CEOs")
```



The mean and median age for CEOs are fairly close with a mean of about 56 and a median of 57. Furthermore, half of CEOs are between 51 and 61 years old.

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

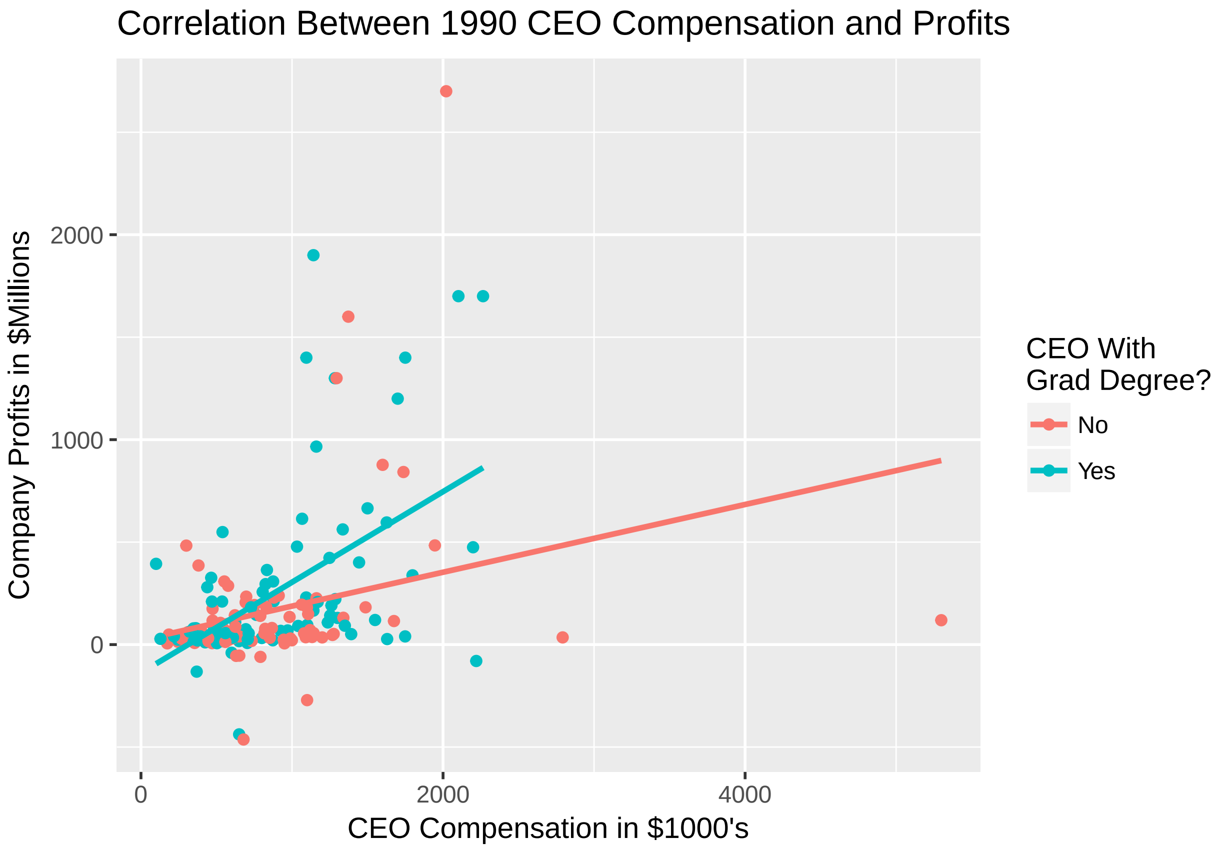
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   51.00   57.00   55.97   61.25   86.00
```

Multivariate Data Analysis

Salary vs Profits

First, we examine the correlation between profits and salary.

```
ggplot(CEO,aes(x = salary, y = profits, color = grad)) +
  geom_point(na.rm = T) +stat_smooth(method = "lm", se = F, na.rm = T) +
   ggtitle("Correlation Between 1990 CEO Compensation and Profits") +
    labs(x = "CEO Compensation in $1000's", y = "Company Profits in $Millions") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



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

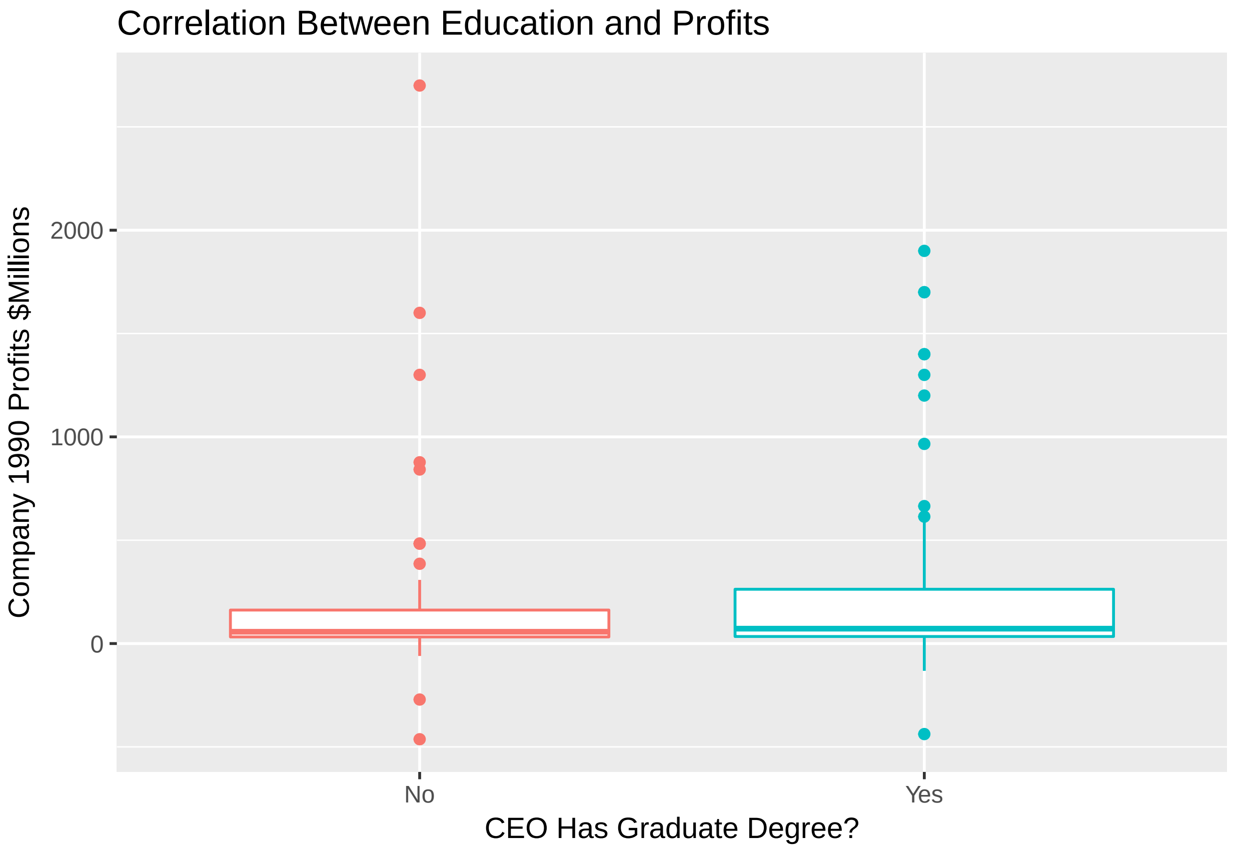
```
## [1] 0.3958275
```

We see that, indeed, salary and profits are relatively well correlated with a correlation coefficient of .399.

Education vs Profit

Next, we'll examine the profits for companies run by those with a graduate level education versus those without it.

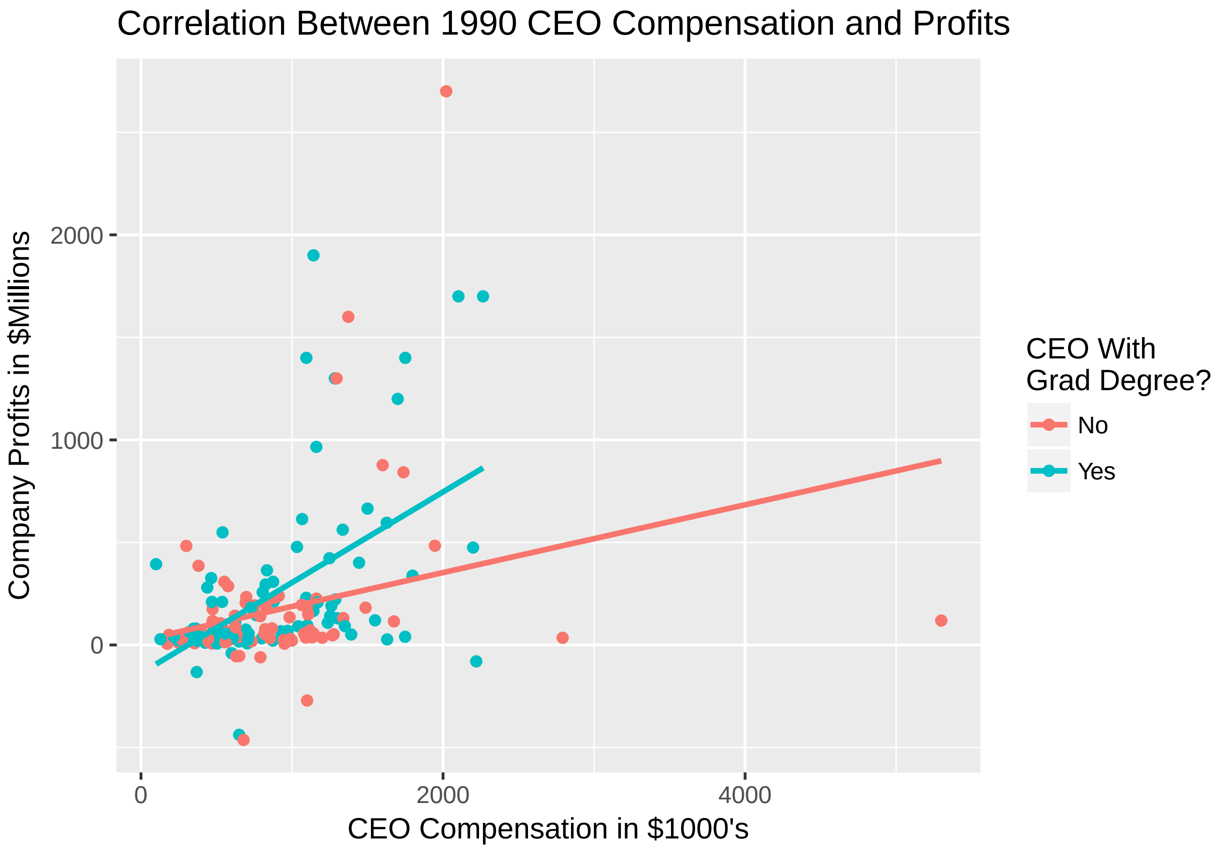
```
ggplot(CEO, aes(x = grad, y = profits, color = grad)) +
  geom_boxplot(na.rm = T) + ggtitle("Correlation Between Education and Profits") +
    labs(x = "CEO Has Graduate Degree?", y = "Company 1990 Profits $Millions") +
     theme(legend.position="none")
```


Correlation Between Education and Profits

Looking at the relationship between education and profits, we can immediately see that both the 25th and 75th percentile for CEOs with graduate degrees is higher than that of the CEOs without a graduate degree.

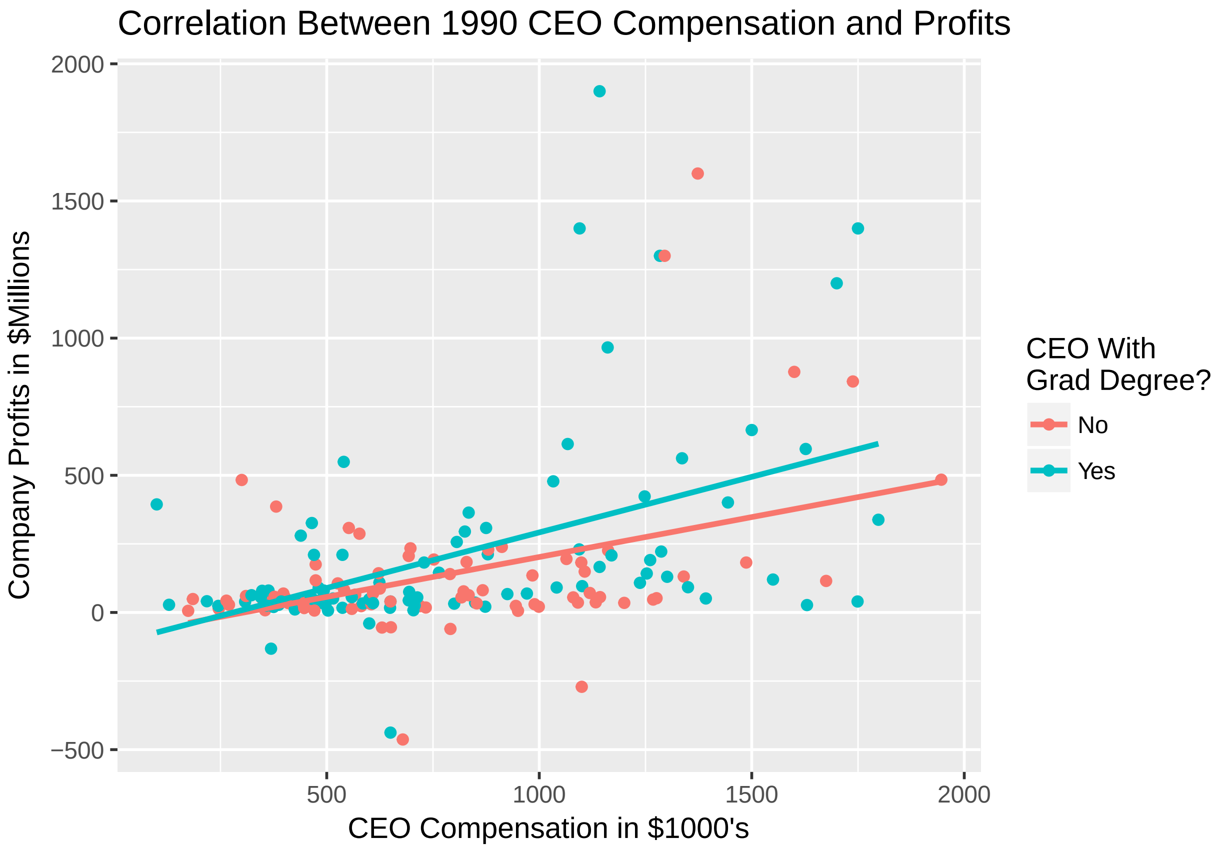
Because of this, we now explore how education might affect the correlation between salary and profits.

```r
ggplot(CEO, aes(x = salary, y = profits, color = grad)) +
  geom_point(na.rm = T) + stat_smooth(method = "lm", se = F, na.rm = T) +
   ggtitle("Correlation Between 1990 CEO Compensation and Profits") +
    labs(x = "CEO Compensation in $1000's", y = "Company Profits in $Millions") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



We see two outliers in the above chart that may be causing the estimation of correlation for non-graduate degree holders to be artificially low, so for this analysis we will remove CEOs with a salary of > $2,000,000. Perhaps the high pay of these CEOs is a sign of more problems within the company and thus why the profits are so low, or perhaps these data points represent companies with a low market value that hired a CEO that they hope will lead to greater growth. Regardless, these data points are outliers that make our analysis less meaningfull.

```
ggplot( subset(CEO, salary < 2000), aes(x = salary, y = profits, color = grad)) +
  geom_point(na.rm = T) +stat_smooth(method = "lm", se = F, na.rm = T) +
   ggtitle("Correlation Between 1990 CEO Compensation and Profits") +
    labs(x = "CEO Compensation in $1000's", y = "Company Profits in $Millions") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```

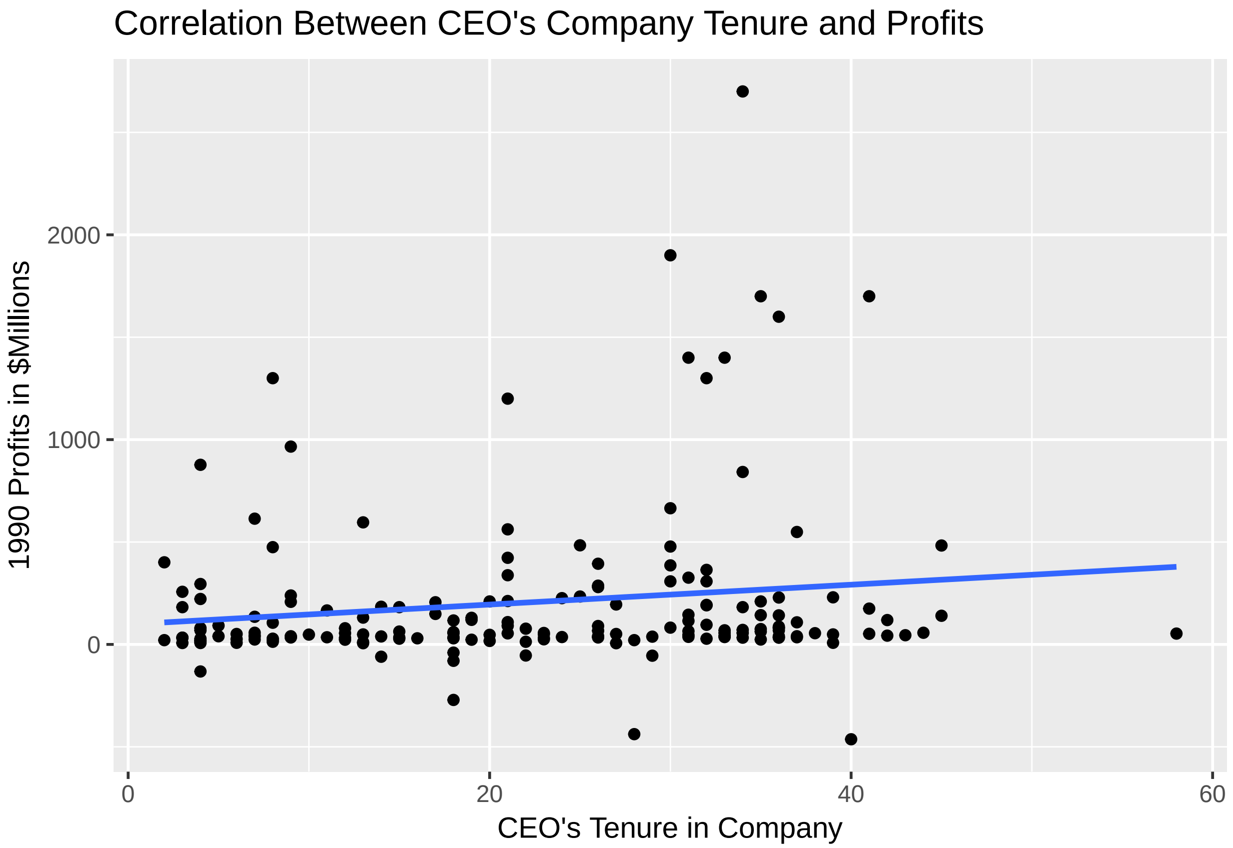


We now see a stronger relationship between salary and profits for non graduate degree holders. Still, the relationship is stronger for those holding a graduate degree, and we can see that, typically, someone with the same salary who holds a graduate degree tends to correlate to higher profits than someone at that same salary level without a graduate degree.

Company Tenure vs Profits

Next, we will examine the correlation between company tenure and profits

```
ggplot(CEO, aes(x = comten, y = profits)) + geom_point(na.rm = T) +
  stat_smooth(method = "lm", se = FALSE, na.rm = T) +
  labs(x = "CEO's Tenure in Company", y = "1990 Profits in $Millions") +
  ggtitle("Correlation Between CEO's Company Tenure and Profits")
```



While most data points are clustered closer to the 0 - 1000 range for all company tenure levels, we see that there are more high profits for companies with CEOs with high company tenure, though the same is true for negative profits.

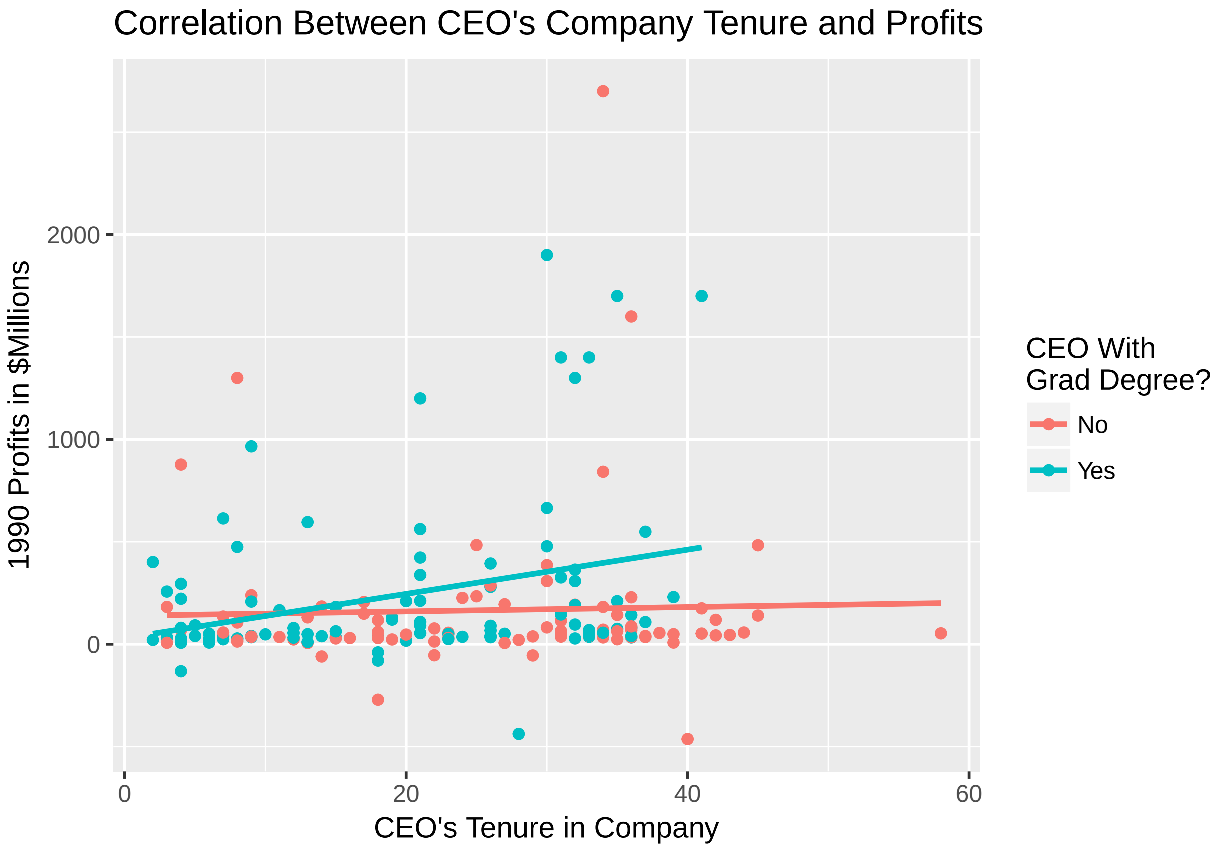
It seems that there might potentially be a weak correlation between company tenure and profits. In fact, the correlation is .1491, confirming that a very weak correlation exists.

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

```
## [1] 0.1491889
```

We now explore how the correlation of company tenure on profits changes based on whether or not the CEO has a graduate degree.

```
ggplot(CEO, aes(x = comten, y = profits, color = grad)) + geom_point(na.rm = T) +
    stat_smooth(method = "lm", se = FALSE, na.rm = T) +
  labs(x = "CEO's Tenure in Company", y = "1990 Profits in $Millions") +
  ggtitle("Correlation Between CEO's Company Tenure and Profits") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```

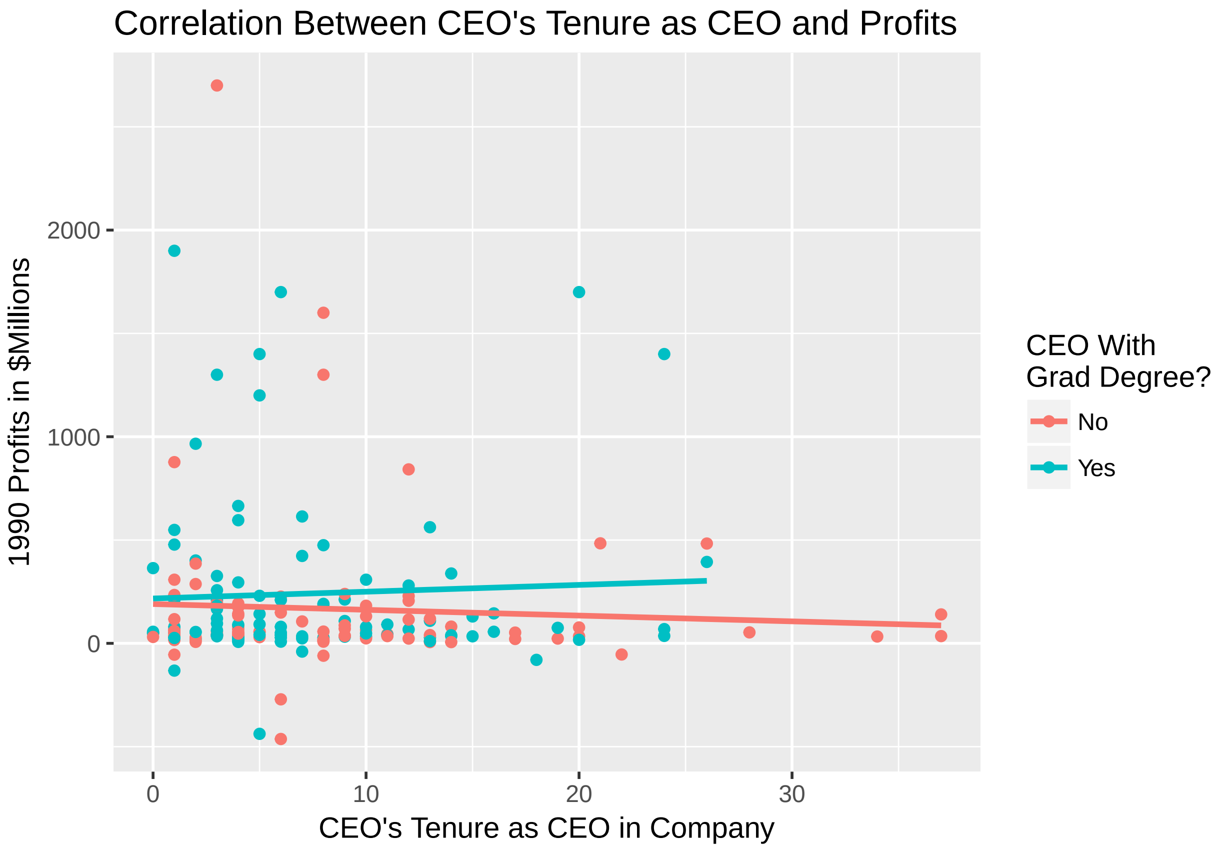


We can see a very interesting trend here. The group of companies with a CEO who received a graduate-level education had a higher correlation between the CEO's company tenure and profits.

Tenure as CEO vs profits

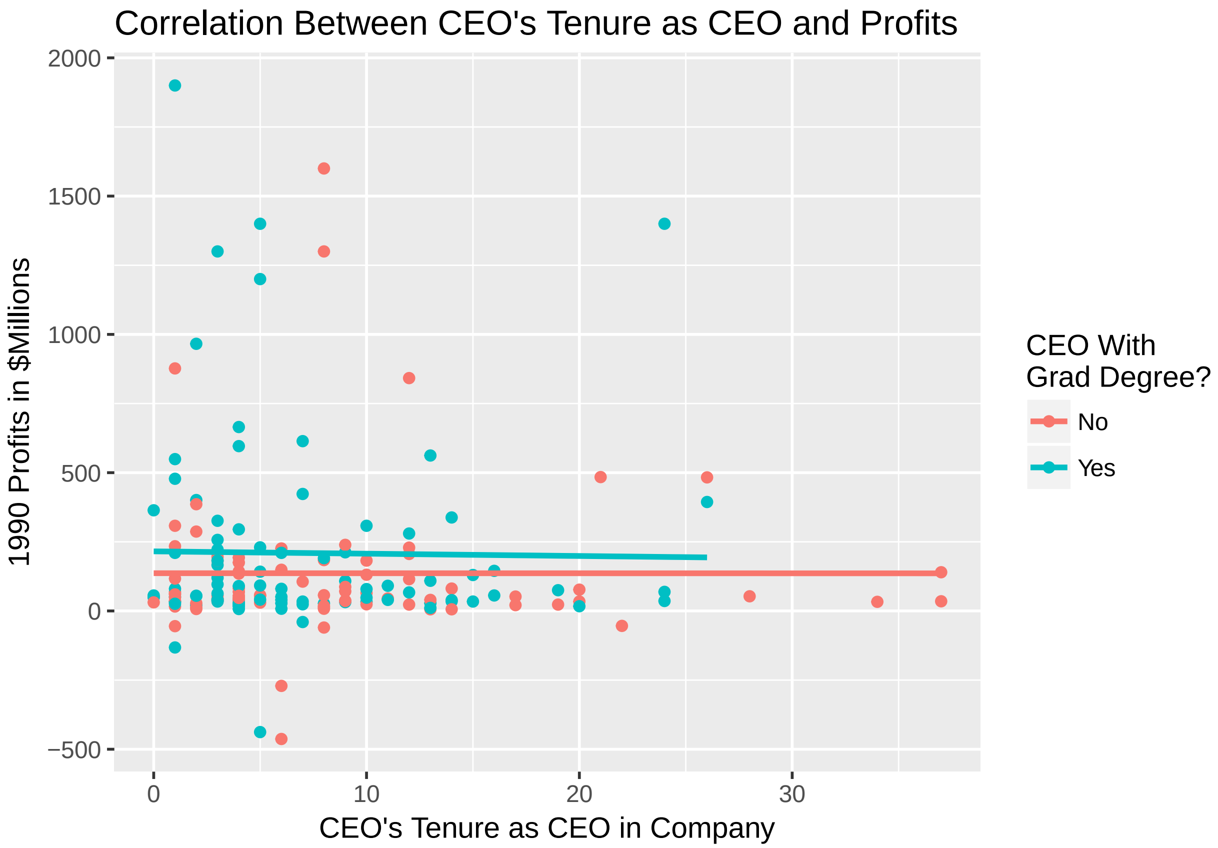
Next, we see the relationship between tenure as CEO vs profits.

```
ggplot(CEO ,aes(x = ceoten, y = profits, color = grad)) +
  geom_point(na.rm = T) + stat_smooth(method = "lm", se = F, na.rm = T) +
  labs(x = "CEO's Tenure as CEO in Company", y = "1990 Profits in $Millions") +
  ggtitle("Correlation Between CEO's Tenure as CEO and Profits") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



We see some odd behavior in the correlation between tenure as CEO and profits, so like before we remove the odd data points with CEOs who had salaries greater than $2,000,000.

```
ggplot(subset(CEO, salary <2000) ,aes(x = ceoten, y = profits, color = grad)) +
  geom_point(na.rm = T) + stat_smooth(method = "lm", se = F, na.rm = T) +
  labs(x = "CEO's Tenure as CEO in Company", y = "1990 Profits in $Millions") +
  ggtitle("Correlation Between CEO's Tenure as CEO and Profits") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



We now see a very interesting story. It seems tenure as a CEO is not correlated with profits, regardless of education level.

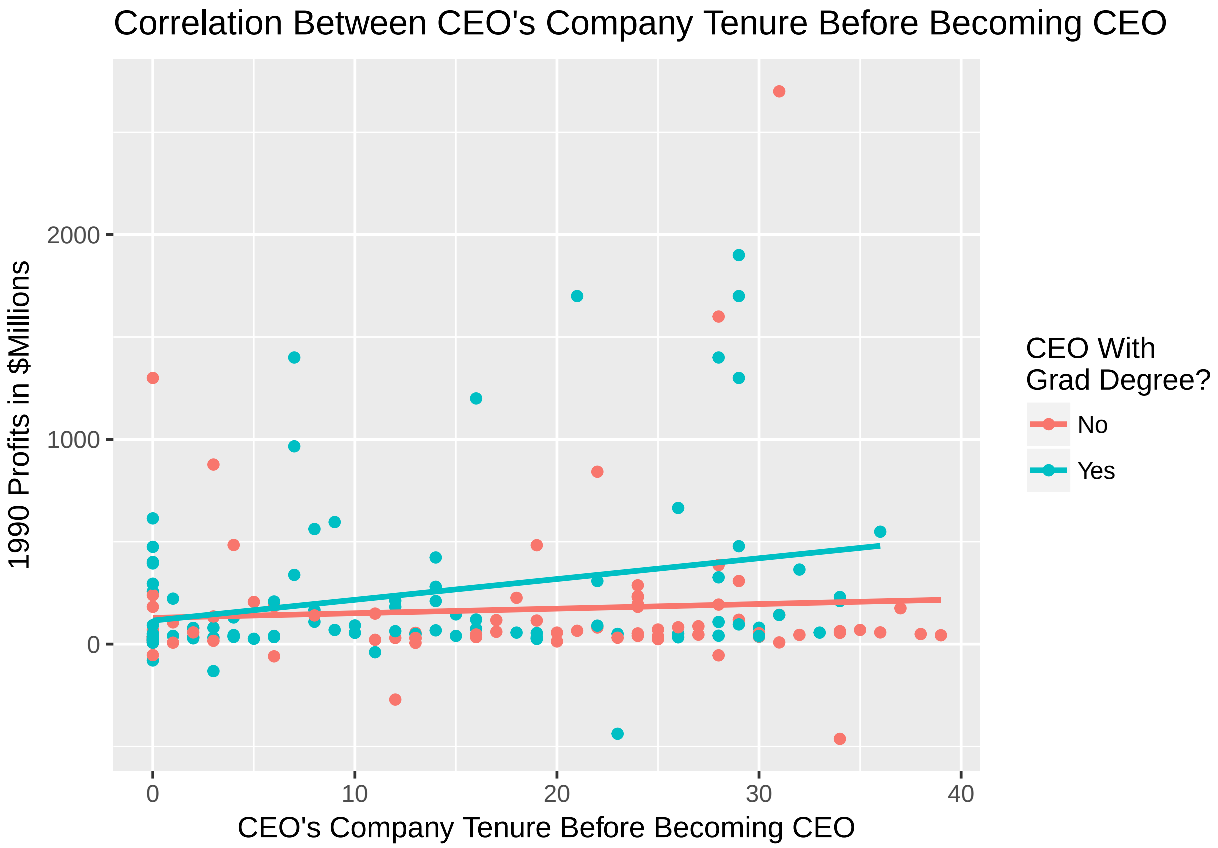
What exactly does this imply? Perhaps the knowledge gathered as a CEO is not as important as the knowledge gathered working in roles other than CEO.

We explore this question further by looking at tenure at the company prior to becoming CEO.

Tenure at Company Prior to Becoming CEO vs Profits

As we saw in the previous sections, there is a correlation between company tenure and profits but not in tenure as a CEO and profits. So we'd like to explore how strongly correlated years at the company before becoming CEO is with profits.

```
CEO$tenure_before_CEO = CEO$comten - CEO$ceoten
ggplot(CEO, aes(x = tenure_before_CEO, y = profits, color = grad)) +
  geom_point(na.rm = T) + stat_smooth(method = "lm", se = F, na.rm = T) +
  labs(x = "CEO's Company Tenure Before Becoming CEO", y = "1990 Profits in $Millions") +
  ggtitle("Correlation Between CEO's Company Tenure Before Becoming CEO") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```

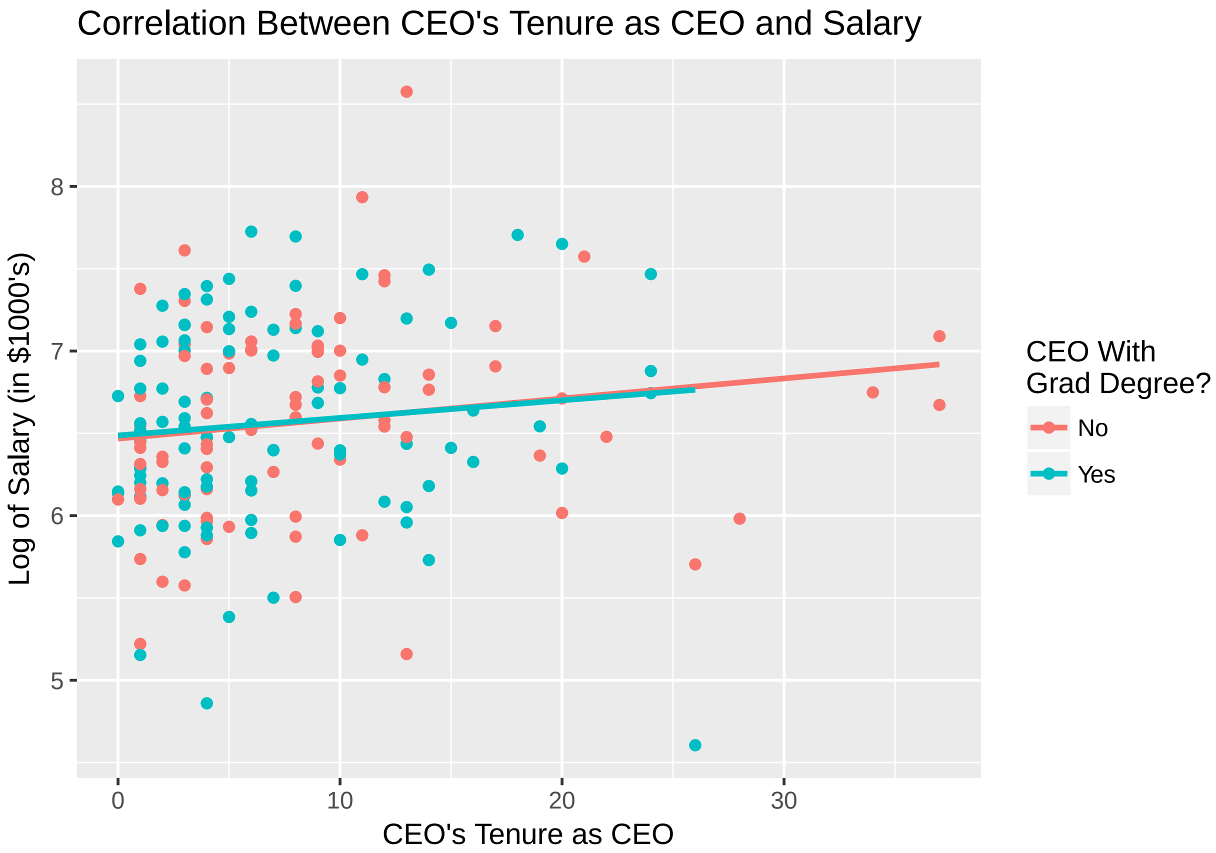


Again we see a stronger correlation between tenure before becoming CEO for those with a graduate degree than those without one. Furthermore, this also seems to further corroborate the idea that perhaps knowing more about the day-to-day details of working at a given company allows a CEO to perform better as a company leader and thus allows the company to earn higher profits.

Tenure as CEO vs Salary

Next we also explore the relationship between CEO and salary. Because of the level of dispersion between tenure as CEO and salary, we perform a log transformation of salary.

```
ggplot(CEO,aes(x = ceoten, y = log(salary), color = grad)) +
  geom_point() + stat_smooth(method = "lm", se = F) +
  labs(x = "CEO's Tenure as CEO", y = "Log of Salary (in $1000's)") +
  ggtitle("Correlation Between CEO's Tenure as CEO and Salary") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



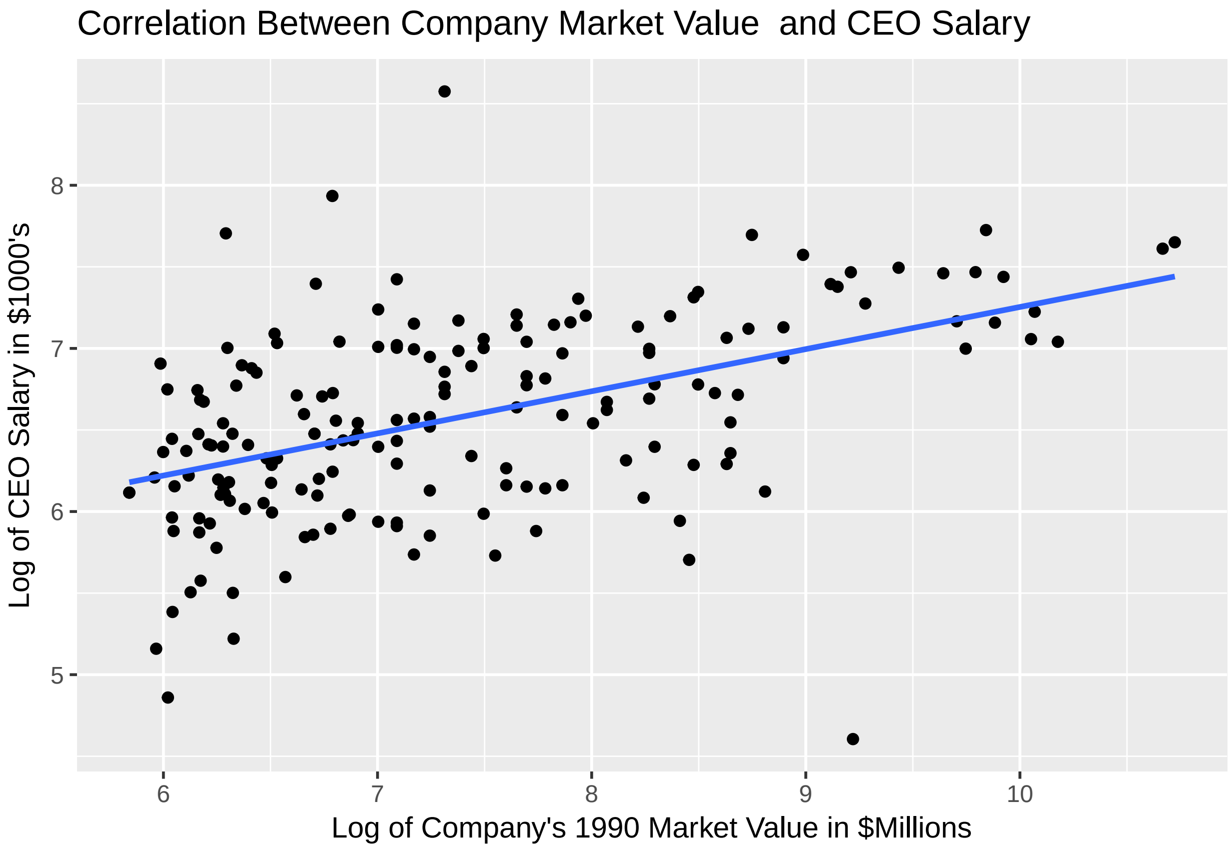
We see that there is a fairly strong linear relationship between tenure as CEO and salary, and that the correlation is roughly similar regardless of education level.

This is expected. The more experience someone has doing a role, we would normally expect them to get paid well. The fact that salaries are so close for CEO's with a graduate degree and those without it is, however, surprising, especially given that profits are typically higher for companies that have a CEO with a graduate level education.

Market Value vs Salary

As we showed earlier, both market value and salary have similar distributions that are strongly right skewed. Both are also always positive, so we will perform log transformations to market value and salary.

```
ggplot(CEO,aes(x = log(mktval), y = log(salary))) +
  geom_point(na.rm = T) + stat_smooth(na.rm = T, method = "lm", se = F)  +
  labs(x = "Log of Company's 1990 Market Value in $Millions", y = "Log of CEO Salary in $1000's") +
  ggtitle("Correlation Between Company Market Value  and CEO Salary")
```



We see a fairly strong correlation between the log of market value and the log of salary, and in fact, with a correlation coefficient of .48, we see that indeed there is a correlation between the two variables.

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

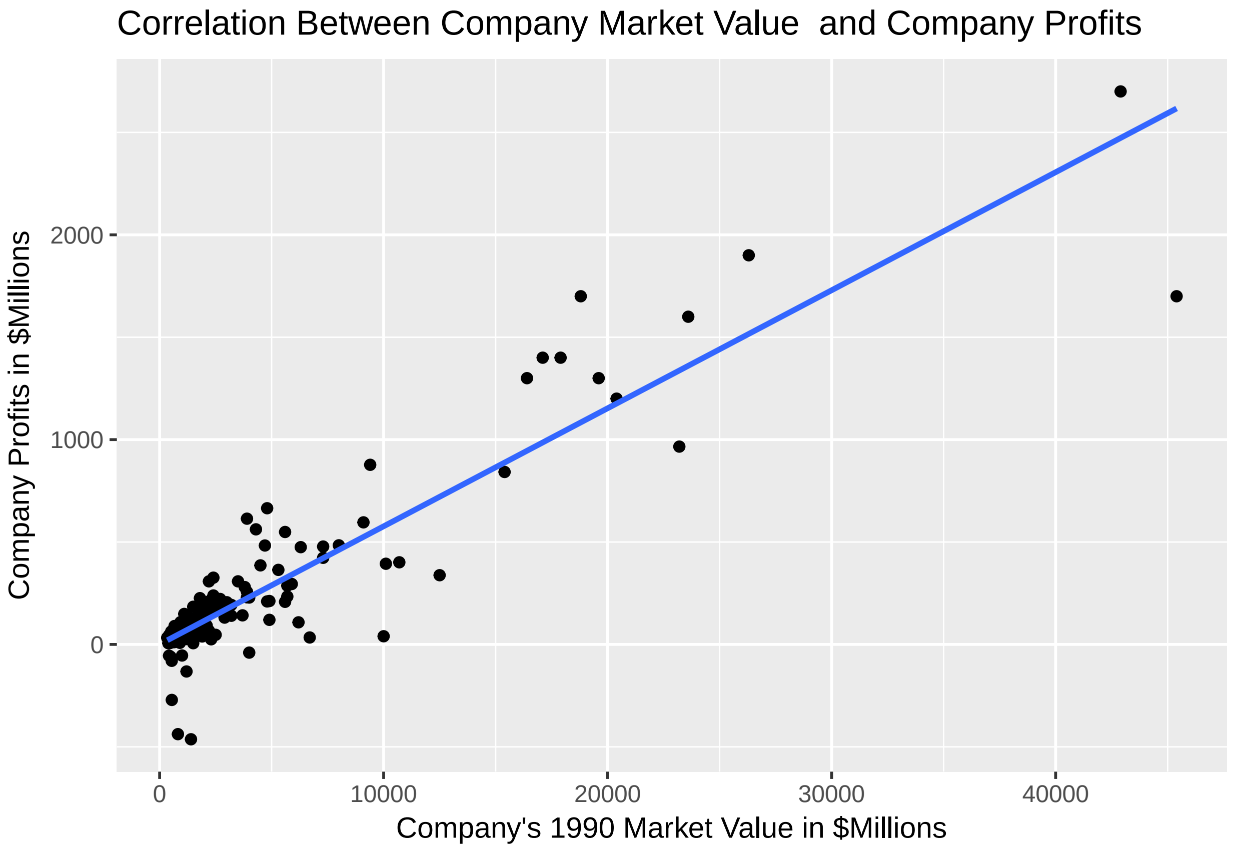
```
## [1] 0.4866164
```

One potential explanation for this correlation is that companies with higher market value also typically have more money than those that have low market value. As such they can afford to pay their employees higher. The CEO, having to oversee, theoretically, a more complicated business will also be able to demand a higher wage.

Market Value vs Profits

Finally, we explore the correlation between market value and profits. Since profits can be negative, no transformation is performed on it, but it is performed on market value.

```
ggplot(CEO,aes(x = mktval, y = profits)) +
  geom_point(na.rm = T) + stat_smooth(na.rm = T, method = "lm", se = F) +
  labs(x = "Company's 1990 Market Value in $Millions", y = "Company Profits in $Millions") +
  ggtitle("Correlation Between Company Market Value  and Company Profits")
```



We see a strong linear relationship between the market value of profits, and indeed with a correlation coefficient of .91, this is the strongest correlation we've seen in this data.

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

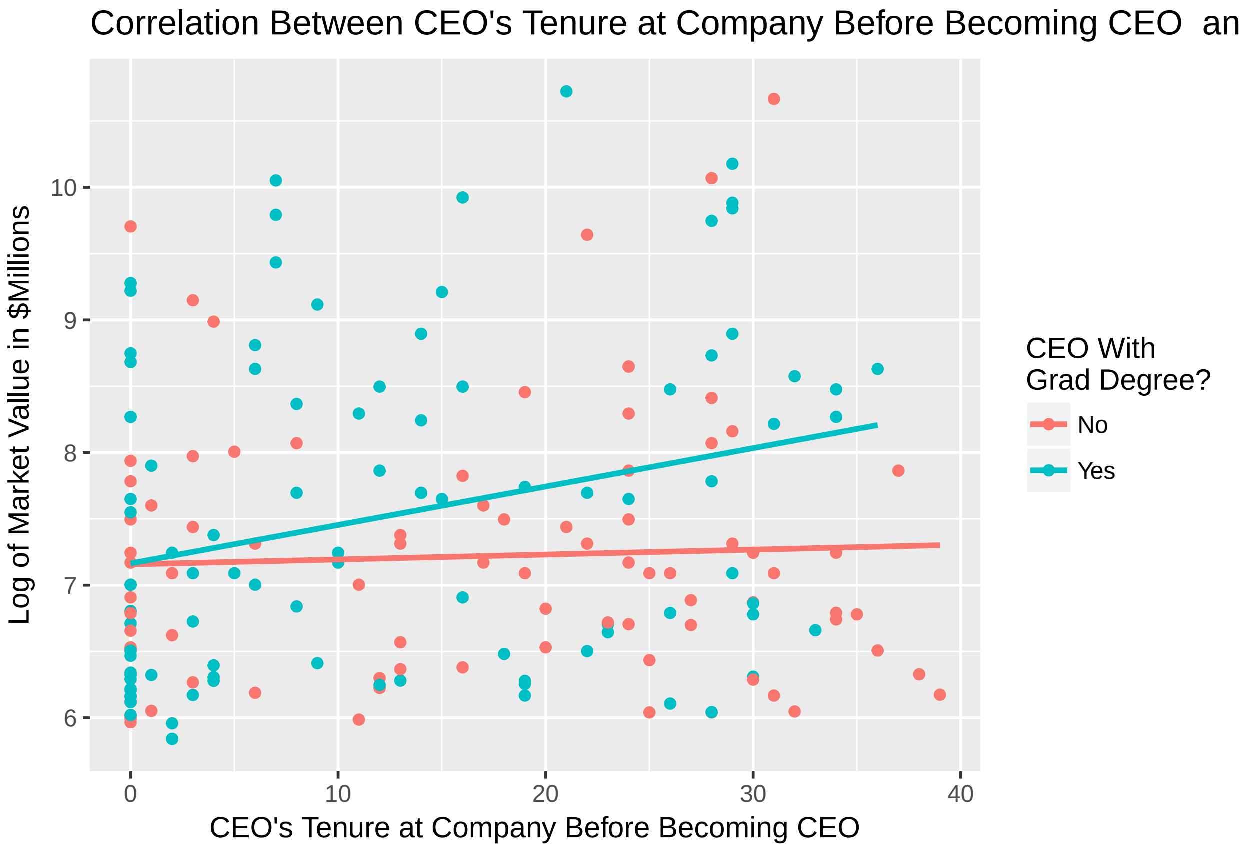
```
## [1] 0.918301
```

Nothing too surprising comes out of this analysis. It makes sense that a company with higher profits also has a higher market value.

Market Value and CEO's Company Tenure Before Becoming CEO

We now explore how a company's value might correlate with a CEO's tenure at a company before he or she becomes a CEO.

```
ggplot(CEO, aes( x = tenure_before_CEO, y = log(mktval), color = grad)) +
  geom_point(na.rm = T) + stat_smooth(method = "lm", se = F, na.rm = T) +
  labs(x = "CEO's Tenure at Company Before Becoming CEO", y = "Log of Market Vallue in $Millions") +
  ggtitle("Correlation Between CEO's Tenure at Company Before Becoming CEO  and Market Value") +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



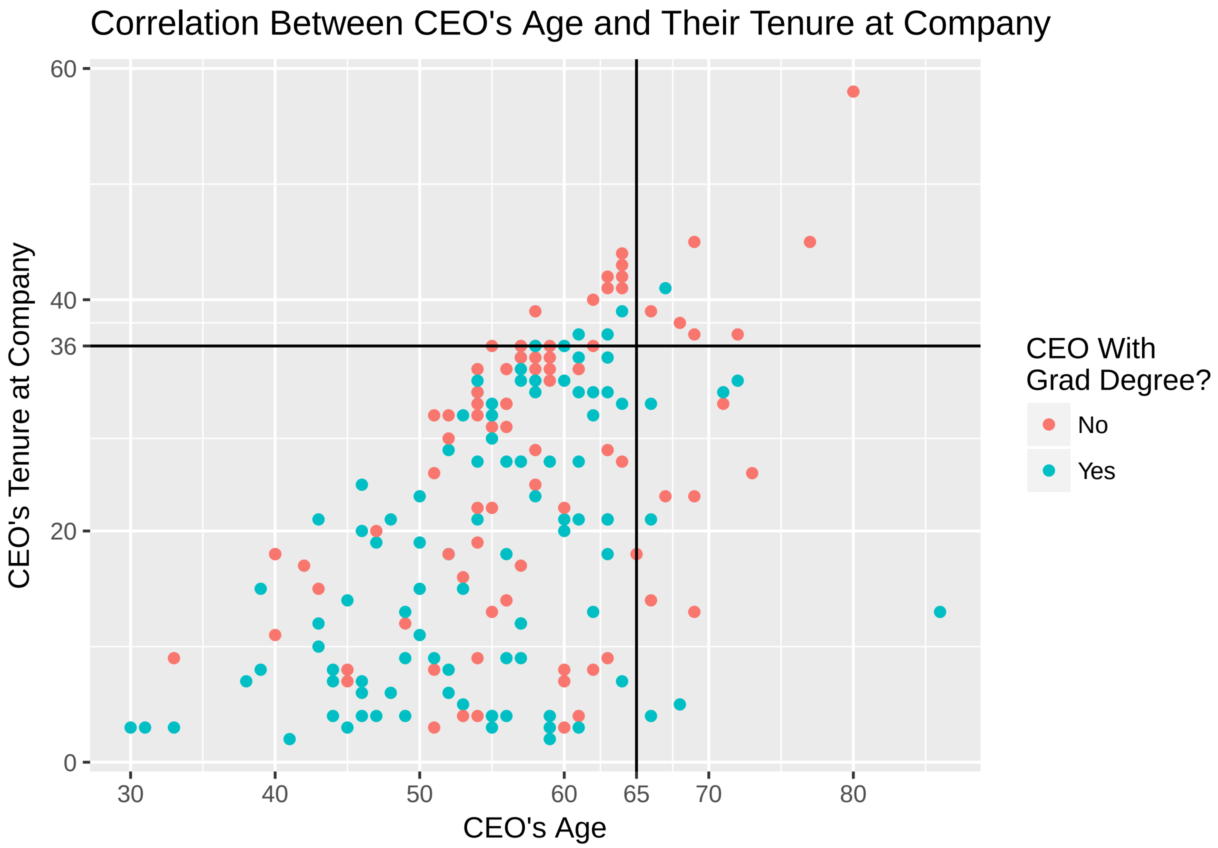
We see here that there is a correlation between the CEO's tenure at a company before he or she becomes a CEO and the market value of the company. However this correlation is a lot less meaningful for CEO's without a graduate degree.

The true meaning of this is difficult to asses. Perhaps large companies simply prefer promoting from within and tend to demand more education from their employees. Alternatively, the knowledge a CEO is able to collect before becoming a CEO allows him or her to understand the impact of his or her decisions.

Age and Company Tenure

Out of curiosity, we wanted to explore why we saw the company tenure of a CEO drop off after 36 and to confirm our theory that this typically correlated with a CEO hitting retirement age, so we plotted age and company tenure against each other

```
ggplot(CEO,aes(y = comten, x = age, color = grad)) +
  geom_point() + geom_vline(xintercept= 65) + geom_hline(yintercept = 36) +
  labs(x = "CEO's Age", y = "CEO's Tenure at Company") +
  ggtitle("Correlation Between CEO's Age and Their Tenure at Company") +
  scale_x_continuous(breaks=c(30,40,50,60,65,70,80)) + scale_y_continuous(breaks=c(0,20,36,40,60)) +
     scale_colour_discrete(name = "CEO With \nGrad Degree?")
```



The upper right square represents CEOs who are older than 65 and who have a company tenure greater than 36. Indeed, we see very few data points in the upper left quadrant, representing CEOs with more than 36 years of company tenure who are younger than 65, so 36 years of company tenure seems to be close to when most CEOs hit their age of retirement.

Furthermore, we also see that CEO's with a graduate degree tend to be older than those who don't have a graduate degree for each level of company tenure. This makes sense. They finish school later, so they don't enter the job market until later. Further more we also see that there are more young CEOs with graduate degrees.

Confounding Effects

Correlation is often confused for causation, so it is important to correctly assess what the data is saying. For instance, we saw that salary is correlated with years of experience, market value of the company, and in deed, profits.

However, we also saw that profits is correlated with company tenure, education, market value, and.

Since multiple correlations exist, it is very possible that for instance, higher education and experience combined allow a CEO to learn how businesses run effectively and thus lead to higher profits, and because the educated and experienced CEO has a history of showing good results, they would demand a higher salary.

Alternatively, it is also possible that simply having a higher salary would motivate a CEO to perform better.

Another alternative could be that companies with high profits and market values tend to have the budget to hire a CEO with a good performance record and pay himor her a very competetive salary.

To understand the true nature of these correlations, it would be necessary to use more advanced modeling techniques.

Conclusion

As explained in the previous section of this report, without a deeper analysis of the data set, it is difficult to conclusively say what the true nature of the relationship between salary and profits other than that they are correlated with each other.

However salary and profits are both correlated with a CEO's company tenure, a company's market value and many more variables.