Analyzing CEO Compensation: Correlations & Predictive Modeling

Anushna Gunda, Inaya Rizvi, Harsh Malik, Sai Bathina

2025-03-11

Table of Contents

[Introduction 1](#_Toc192556613)

[Research Questions 2](#_Toc192556614)

[Data and Methods 2](#_Toc192556615)

[Overview of Dataset 2](#_Toc192556616)

[Data Import & Preprocessing 3](#_Toc192556617)

[Exploratory Analysis 8](#_Toc192556618)

[Summary Statistics 8](#_Toc192556619)

[Overview of CEO Compensation by Location 9](#_Toc192556620)

[We will create a correlation matrix to check relationships between independent variables (avoid multicollinearity) and assess their correlation with our dependent variable (CEO compensation) 15](#_Toc192556621)

[Visualizations 16](#_Toc192556622)

[Regression Analysis 21](#_Toc192556623)

[Multiple Linear Regression: CEO Compensation ~ Financial Metrics 21](#_Toc192556624)

[Scaling Data to standardize 21](#_Toc192556625)

[Results 26](#_Toc192556626)

[Discussion 26](#_Toc192556627)

# Introduction

CEO compensation is a major topic in corporate governance and financial research, sparking debates about fairness, performance incentives, and economic inequality. Executives’ pay structures typically combine salary, bonuses, stock options, and long-term incentives, aiming to align CEO and shareholder interests. However, a critical question remains: is CEO pay primarily driven by company performance, or do other structural and market factors play a bigger role? Understanding these dynamics is essential for investors, policymakers, and corporate boards to ensure transparency, accountability, and effective compensation strategies.

Previous research has examined the key drivers of executive pay. Anderson, Banker, and Ravindran (2000) found that firm size and growth opportunities significantly impact CEO salaries in the IT sector. Banerjee (2022) highlighted global salary discrepancies, while Bouteska, Sharif, and Zoghbi (2024) linked CEO pay to both firm performance and risk-taking behavior. Elsayed and Elbardan (2018) found that high-powered incentive structures often lead to earnings manipulation, further complicating the discussion. Despite these insights, there is no widely accepted model that accurately predicts CEO compensation across industries.

In this study, we aim to build on these findings using a dataset sourced from PitchBook. Our analysis is guided by two key questions:

1. What financial and market factors most strongly influence CEO compensation?
2. Can we develop a predictive model to estimate CEO total compensation based on these indicators?

To answer these, we apply statistical techniques including correlation analysis, multiple linear regression, and data visualization to identify the strongest predictors of CEO pay. By understanding these relationships, our findings can help investors evaluate whether compensation levels are justified, assist corporate boards in designing performance-driven pay structures, and provide regulators with insights into potential governance risks. This study contributes to the broader conversation on executive compensation by using data-driven methods to examine how financial and market indicators shape CEO pay.

# Research Questions

* What company financial and market factors influence CEO compensation?
* Can we create a predictive model to predict CEO total compensation?

# Data and Methods

**Source:** Pitchbook UW

## Overview of Dataset

Using Pitchbook, we custom-formatted a dataset that includes **CEO total compensation** as the dependent variable and the following independent variables:

* **Company Financials:** Total Revenue, Revenue per Employee
* **Profitability Metrics:** ROIC, ROA, Net Profit Margin, EBITDA Margin, Revenue Growth
* **Debt & Leverage:** Net Debt, Debt to Capital, Total Assets
* **Market Indicators:** Price % Change YTD, EV
* **Categorical Factors:** HQ State/Province, Primary Industry Group, CEO

## Data Import & Preprocessing

Includes formatting column names and negative values (ceo\_data), and handling null values by replacing null with medians (quantitative) or mode (qualitative) (imputed\_data)

As a result of pre-processing the data, there are 2 datasets that we use for analysis/exploration: *ceo\_data* = clean but with NA values for missing data *imputed\_data* = clean with NA values replaced with median or mode (depending on data type)

# Loading Required Libraries  
library(tidyverse) # Data manipulation  
library(mice) # Missing data imputation  
library(dplyr)  
  
# Loading the Dataset  
raw\_data <- read.csv("ceo\_comp\_data.csv", na.strings = c("", "NA"))  
  
# Fixing Column Names  
ceo\_data <- raw\_data %>%  
 rename(  
 total\_revenue = "Total.Revenue..FY.",  
 revenue\_per\_employee= "Revenue.per.Employee..FY.",   
 roa = "ROA..Return.on.Asset...FY.",  
 roic = "ROIC..Return.on.Invested.Capital...FY.",  
 net\_profit\_margin = "Net.Profit.Margin..FY.",  
 ebitda\_margin = "EBITDA.Margin..FY.",  
 revenue\_growth = "Revenue...Growth..FY.",  
 net\_debt = "Net.Debt..FY.",  
 debt\_to\_capital = "Debt.to.Capital..FY.",  
 total\_assets = "Total.Assets..FY.",  
 price\_change\_ytd = "Price...Change.YTD",  
 ev = "EV..FY.",  
 hq = "HQ.State.Province",  
 primary\_industry\_group = "Primary.Industry.Group",  
 ceo = "CEO",  
 ceo\_total\_comp\_mils = "CEO.Total.Compensation..in.millions.",  
 employees = "Employees",  
 companies = "Companies"  
 )  
  
# Viewing the new column names  
colnames(ceo\_data)

## [1] "companies" "employees" "total\_revenue"   
## [4] "revenue\_per\_employee" "roa" "roic"   
## [7] "net\_profit\_margin" "ebitda\_margin" "revenue\_growth"   
## [10] "net\_debt" "debt\_to\_capital" "total\_assets"   
## [13] "price\_change\_ytd" "ev" "hq"   
## [16] "primary\_industry\_group" "ceo" "ceo\_total\_comp\_mils"

# Function to clean and convert numeric columns with commas & parentheses  
convert\_numeric\_column <- function(column) {  
 column <- as.character(column) # Ensure it's a character string first  
   
 # Remove any leading/trailing spaces  
 column <- trimws(column)  
   
 # Convert negative values (e.g., "(2,000)" → "-2000")  
 column <- gsub("[(),]", "", column) # Remove parentheses and commas  
 column <- ifelse(grepl("^\\(.\*\\)$", column), paste0("-", column), column)   
   
 # Convert to numeric  
 column <- as.numeric(column)  
   
 return(column)  
}  
  
# Apply conversion to `total\_revenue` & `total\_assets`  
ceo\_data$total\_revenue <- convert\_numeric\_column(ceo\_data$total\_revenue)  
ceo\_data$total\_assets <- convert\_numeric\_column(ceo\_data$total\_assets)  
ceo\_data$employees <- convert\_numeric\_column(ceo\_data$employees)  
  
# Check if conversion worked  
str(ceo\_data$total\_revenue)

## num [1:367] 6.41e+07 6.32e+06 1.94e+07 6.05e+07 2.12e+08 ...

str(ceo\_data$total\_assets)

## num [1:367] 5.12e+07 6.67e+06 1.79e+07 1.27e+08 4.12e+08 ...

numeric\_cols <- c("roa", "roic", "net\_profit\_margin", "ebitda\_margin", "revenue\_growth",   
 "net\_debt", "price\_change\_ytd", "ev", "revenue\_per\_employee",  
 "debt\_to\_capital", "ceo\_total\_comp\_mils")  
  
# Defining function to properly format negative numeric values   
convert\_negatives\_column <- function(column) {  
 column <- as.character(column) # Convert to character  
   
 for (i in seq\_along(column)) {   
 if (!is.na(column[i]) && column[i] != "") { # Skip NA or empty values  
 column[i] <- trimws(column[i]) # Remove any leading/trailing spaces  
   
 if (grepl("^\\(.\*\\)$", column[i])) { # Check if value has parentheses  
 column[i] <- gsub("[(),]", "", column[i]) # Remove parentheses & commas  
 column[i] <- as.numeric(column[i]) \* -1 # Convert to negative  
 } else {  
 column[i] <- gsub(",", "", column[i]) # Remove commas from numbers  
 column[i] <- as.numeric(column[i]) # Convert normally  
 }  
   
 # If conversion fails, set it to NA explicitly (to catch errors)  
 if (is.na(column[i])) {  
 column[i] <- NA  
 }  
 }  
 }  
   
 return(as.numeric(column)) # Ensure the final output is numeric  
}  
  
  
# Apply fix to all numeric columns  
for (col in numeric\_cols) {  
 ceo\_data[[col]] <- convert\_negatives\_column(ceo\_data[[col]])  
}  
  
  
# Create a Copy for Imputation  
imputed\_data <- ceo\_data  
  
# Defining function to impute numeric columns with median  
impute\_median <- function(column) {  
 column[is.na(column)] <- median(column, na.rm = TRUE)  
 return(column)  
}  
  
# Imputing numeric columns  
imputed\_data[numeric\_cols] <- lapply(imputed\_data[numeric\_cols], impute\_median)  
  
  
# Defining function to impute categorical columns with mode  
mode\_impute <- function(column) {  
 column[is.na(column)] <- names(sort(table(column), decreasing = TRUE))[1]  
 return(column)  
}  
  
# Imputing categorical columns   
categorical\_cols <- c("primary\_industry\_group", "hq")   
imputed\_data[categorical\_cols] <- lapply(imputed\_data[categorical\_cols], mode\_impute)  
  
  
# Ensuring that all numeric columns are actually numeric type for both   
ceo\_data[numeric\_cols] <- lapply(ceo\_data[numeric\_cols], function(x) as.numeric(as.character(x)))  
imputed\_data[numeric\_cols] <- lapply(imputed\_data[numeric\_cols], function(x) as.numeric(as.character(x)))  
  
  
# Check the Cleaned Data  
glimpse(ceo\_data)

## Rows: 367  
## Columns: 18  
## $ companies <chr> "Accenture (NYS: ACN)", "Concentrix (IT Service…  
## $ employees <dbl> 774000, 450000, 336800, 270300, 228000, 183323,…  
## $ total\_revenue <dbl> 64111745, 6324473, 19428000, 60530000, 21191500…  
## $ revenue\_per\_employee <dbl> 87464.86, 20077.69, 54680.55, 194442.66, 958891…  
## $ roa <dbl> 14.04, 2.06, 11.72, 4.44, 18.96, 23.78, 26.93, …  
## $ roic <dbl> 24.47, 10.34, 16.60, 3.50, 29.16, 20.65, 56.03,…  
## $ net\_profit\_margin <dbl> 10.92, 6.89, 11.79, 2.95, 34.15, 21.20, 25.31, …  
## $ ebitda\_margin <dbl> 16.08, 15.57, 18.25, 11.58, 48.03, 29.34, 32.68…  
## $ revenue\_growth <dbl> 4.09, 13.20, 4.98, 5.54, 6.88, 9.78, -2.80, 17.…  
## $ net\_debt <dbl> -5900573, 2078906, -967000, 45275000, -51291000…  
## $ debt\_to\_capital <dbl> 0.12, 0.45, 0.11, 0.71, 0.23, 0.10, 0.64, 0.99,…  
## $ total\_assets <dbl> 51245305, 6669768, 17852000, 127243000, 4119760…  
## $ price\_change\_ytd <dbl> 31.51, -26.25, 32.07, 16.08, 56.80, 58.32, 48.1…  
## $ ev <dbl> 199518392, 8554569, 28180167, 171710363, 248818…  
## $ hq <chr> NA, "California", "New Jersey", "New York", "Wa…  
## $ primary\_industry\_group <chr> "IT Services", "Commercial Services", "IT Servi…  
## $ ceo <chr> "Julie Sweet", "Christopher Caldwell", "Ravi Ku…  
## $ ceo\_total\_comp\_mils <dbl> 31.55, 10.24, 22.56, 20.40, 48.51, 8.80, 63.21,…

# Check the Imputed Dataset  
glimpse(imputed\_data)

## Rows: 367  
## Columns: 18  
## $ companies <chr> "Accenture (NYS: ACN)", "Concentrix (IT Service…  
## $ employees <dbl> 774000, 450000, 336800, 270300, 228000, 183323,…  
## $ total\_revenue <dbl> 64111745, 6324473, 19428000, 60530000, 21191500…  
## $ revenue\_per\_employee <dbl> 87464.86, 20077.69, 54680.55, 194442.66, 958891…  
## $ roa <dbl> 14.04, 2.06, 11.72, 4.44, 18.96, 23.78, 26.93, …  
## $ roic <dbl> 24.47, 10.34, 16.60, 3.50, 29.16, 20.65, 56.03,…  
## $ net\_profit\_margin <dbl> 10.92, 6.89, 11.79, 2.95, 34.15, 21.20, 25.31, …  
## $ ebitda\_margin <dbl> 16.08, 15.57, 18.25, 11.58, 48.03, 29.34, 32.68…  
## $ revenue\_growth <dbl> 4.09, 13.20, 4.98, 5.54, 6.88, 9.78, -2.80, 17.…  
## $ net\_debt <dbl> -5900573, 2078906, -967000, 45275000, -51291000…  
## $ debt\_to\_capital <dbl> 0.12, 0.45, 0.11, 0.71, 0.23, 0.10, 0.64, 0.99,…  
## $ total\_assets <dbl> 51245305, 6669768, 17852000, 127243000, 4119760…  
## $ price\_change\_ytd <dbl> 31.51, -26.25, 32.07, 16.08, 56.80, 58.32, 48.1…  
## $ ev <dbl> 199518392, 8554569, 28180167, 171710363, 248818…  
## $ hq <chr> "California", "California", "New Jersey", "New …  
## $ primary\_industry\_group <chr> "IT Services", "Commercial Services", "IT Servi…  
## $ ceo <chr> "Julie Sweet", "Christopher Caldwell", "Ravi Ku…  
## $ ceo\_total\_comp\_mils <dbl> 31.55, 10.24, 22.56, 20.40, 48.51, 8.80, 63.21,…

# Optional - Save Cleaned Datasets  
# write.csv(ceo\_data, "ceo\_comp\_data\_raw.csv", row.names = FALSE)  
# write.csv(imputed\_data, "ceo\_comp\_data\_imputed.csv", row.names = FALSE)

## Exploratory Analysis

### Summary Statistics

#summary(imputed\_data)  
  
#Mean and standard deviation for numeric columns  
data\_stats <- data.frame(  
 Mean = sapply(ceo\_data[numeric\_cols], mean, na.rm = TRUE),  
 SD = sapply(ceo\_data[numeric\_cols], sd, na.rm = TRUE),  
 Min = sapply(ceo\_data[numeric\_cols], min, na.rm = TRUE),  
 Max = sapply(ceo\_data[numeric\_cols], max, na.rm = TRUE),  
 Median = sapply(ceo\_data[numeric\_cols], median, na.rm = TRUE),  
 Q1 = sapply(ceo\_data[numeric\_cols], quantile, probs = 0.25, na.rm = TRUE),  
 Q3 = sapply(ceo\_data[numeric\_cols], quantile, probs = 0.75, na.rm = TRUE),  
 Missing = sapply(ceo\_data[numeric\_cols], function(x) sum(is.na(x)))  
)  
  
#Display the detailed statistics  
print(data\_stats)

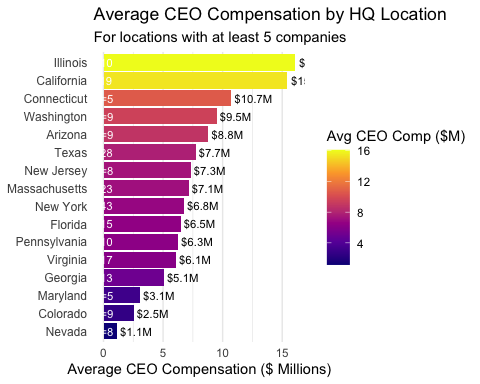
## Mean SD Min Max  
## roa -6.594383e+01 5.875306e+02 -10351.93 1.254800e+02  
## roic -2.143456e+01 9.476776e+01 -968.67 6.864900e+02  
## net\_profit\_margin -4.376448e+03 6.441773e+04 -1209341.68 7.973000e+01  
## ebitda\_margin -4.042229e+03 6.087511e+04 -1139048.64 7.188000e+01  
## revenue\_growth 3.527816e+01 1.341289e+02 -99.92 1.824170e+03  
## net\_debt 7.330865e+05 8.185035e+06 -84083000.00 8.029400e+07  
## price\_change\_ytd 2.665195e+01 7.248749e+01 -98.01 4.780400e+02  
## ev 3.643968e+07 2.117768e+08 -66148.00 2.723535e+09  
## revenue\_per\_employee 3.915228e+05 4.048536e+05 0.00 3.657775e+06  
## debt\_to\_capital 4.523014e-01 2.829120e+00 -10.00 4.755000e+01  
## ceo\_total\_comp\_mils 9.922186e+00 1.703306e+01 0.00 1.618300e+02  
## Median Q1 Q3 Missing  
## roa -0.380 -16.5050 6.785000e+00 4  
## roic -4.355 -24.1625 8.212500e+00 27  
## net\_profit\_margin -7.740 -38.5950 7.035000e+00 4  
## ebitda\_margin 1.550 -30.0025 1.795000e+01 9  
## revenue\_growth 15.750 4.3850 2.905000e+01 9  
## net\_debt 0.000 -147804.0000 3.510530e+05 6  
## price\_change\_ytd 18.810 -12.6600 6.059000e+01 14  
## ev 1996585.000 225707.5000 1.068062e+07 20  
## revenue\_per\_employee 290008.710 191413.0400 4.452260e+05 26  
## debt\_to\_capital 0.320 0.0600 6.200000e-01 2  
## ceo\_total\_comp\_mils 4.795 1.0750 1.219250e+01 1

### Overview of CEO Compensation by Location

# Load required libraries  
library(ggplot2)  
library(dplyr)  
library(viridis)  
library(stringr)  
library(ggrepel)  
  
# Create location summary statistics  
location\_summary <- imputed\_data %>%  
 filter(!is.na(hq)) %>%  
 group\_by(hq) %>%  
 summarize(  
 avg\_ceo\_comp = mean(ceo\_total\_comp\_mils, na.rm = TRUE),  
 median\_ceo\_comp = median(ceo\_total\_comp\_mils, na.rm = TRUE),  
 count = n(),  
 avg\_ev = mean(ev/1e6, na.rm = TRUE),  
 avg\_employees = mean(employees, na.rm = TRUE)  
 ) %>%  
 # Filter to include only locations with at least 5 companies  
 filter(count >= 5) %>%  
 arrange(desc(avg\_ceo\_comp))  
  
# Print top locations by CEO compensation  
print(location\_summary)

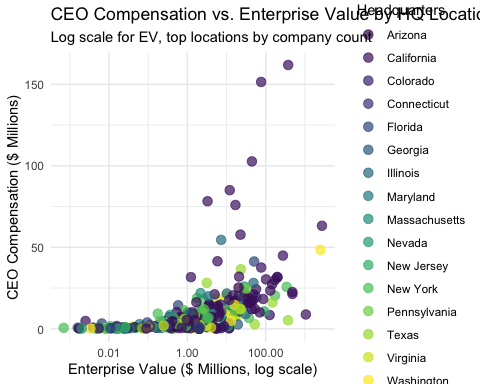
## # A tibble: 16 × 6  
## hq avg\_ceo\_comp median\_ceo\_comp count avg\_ev avg\_employees  
## <chr> <dbl> <dbl> <int> <dbl> <dbl>  
## 1 Illinois 16.1 10.1 10 12.6 6855.  
## 2 California 15.4 7.21 119 63.6 22738.  
## 3 Connecticut 10.7 12.8 5 11.0 13111   
## 4 Washington 9.50 4.27 9 281. 27307.  
## 5 Arizona 8.81 4.22 9 13.1 6526.  
## 6 Texas 7.73 4.95 28 24.0 17954.  
## 7 New Jersey 7.32 5.48 8 15.8 58825.  
## 8 Massachusetts 7.15 5.55 23 5.49 3116   
## 9 New York 6.76 4.47 43 16.4 18511.  
## 10 Florida 6.47 1.66 15 7.89 5016.  
## 11 Pennsylvania 6.26 4.46 10 7.40 9821.  
## 12 Virginia 6.09 2.55 17 5.54 15284.  
## 13 Georgia 5.06 2.76 13 6.39 6369.  
## 14 Maryland 3.05 1.94 5 1.31 663.  
## 15 Colorado 2.54 1.73 9 2.14 7458.  
## 16 Nevada 1.15 0.475 8 0.0538 306.

# Create a bar chart showing average CEO compensation by location  
p\_location\_bar <- ggplot(location\_summary,   
 aes(x = reorder(hq, avg\_ceo\_comp),   
 y = avg\_ceo\_comp,   
 fill = avg\_ceo\_comp)) +  
 geom\_col() +  
 geom\_text(aes(label = sprintf("$%.1fM", avg\_ceo\_comp)),   
 hjust = -0.1, size = 3) +  
 geom\_text(aes(label = sprintf("n=%d", count)),  
 y = 1, hjust = 1.1, size = 3, color = "white") +  
 coord\_flip() +  
 scale\_fill\_viridis\_c(option = "plasma") +  
 theme\_minimal() +  
 labs(  
 title = "Average CEO Compensation by HQ Location",  
 subtitle = "For locations with at least 5 companies",  
 x = NULL,  
 y = "Average CEO Compensation ($ Millions)",  
 fill = "Avg CEO Comp ($M)"  
 ) +  
 theme(  
 panel.grid.major.y = element\_blank(),  
 axis.text.y = element\_text(size = 9)  
 )  
  
print(p\_location\_bar)



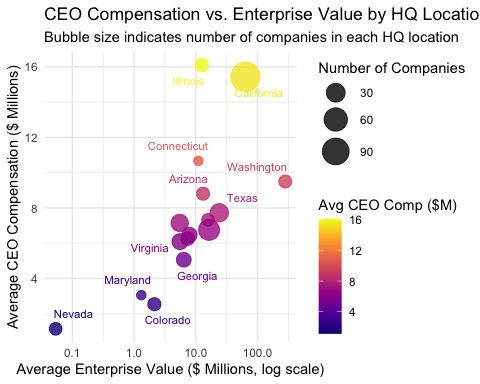
The concentration of businesses in a given area is probably not the only factor influencing CEO compensation; other criteria, such as industry type, firm size, profitability, and market value, are probably more important. Although adding headquarters location as a categorical variable to the regression model might still increase prediction accuracy, the fact that it did not correlate with the number of companies suggests that CEO salary may be more strongly impacted by general market and financial conditions. Based on our research, we compiled a dataset of variables that we hypothesize would correlate with CEO compensation to explore.

# Create a scatter plot comparing CEO compensation vs. enterprise value by location  
p\_location\_scatter <- ggplot(imputed\_data %>%   
 filter(hq %in% location\_summary$hq),   
 aes(x = ev/1e6,   
 y = ceo\_total\_comp\_mils,  
 color = hq)) +  
 geom\_point(alpha = 0.7, size = 3) +  
 scale\_x\_log10(labels = scales::comma\_format()) +  
 scale\_color\_viridis\_d() +  
 theme\_minimal() +  
 labs(  
 title = "CEO Compensation vs. Enterprise Value by HQ Location",  
 subtitle = "Log scale for EV, top locations by company count",  
 x = "Enterprise Value ($ Millions, log scale)",  
 y = "CEO Compensation ($ Millions)",  
 color = "Headquarters"  
 ) +  
 theme(legend.position = "right")  
  
print(p\_location\_scatter)



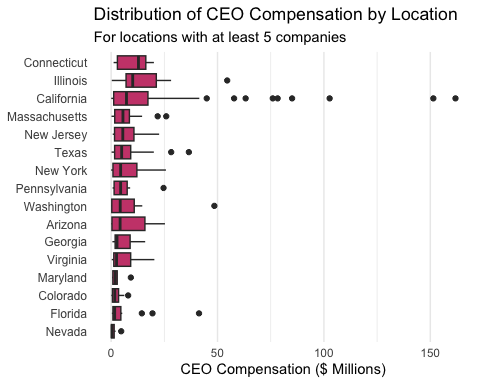
With notable fluctuation and outliers, the scatterplot demonstrates a slight correlation between CEO compensation and enterprise value. Although CEO compensation tends to increase with business valuation, the inconsistent data indicates that other factors, such as profitability, industry type, and market conditions, might have a greater impact. This suggests that CEO compensation cannot be accurately predicted by company value alone.

# Create a location comparison visualization  
p\_location\_comparison <- ggplot(location\_summary,   
 aes(x = avg\_ev,   
 y = avg\_ceo\_comp,   
 size = count,  
 color = avg\_ceo\_comp)) +  
 geom\_point(alpha = 0.8) +  
 geom\_text\_repel(aes(label = hq),   
 size = 3,  
 box.padding = 0.5,  
 point.padding = 0.5,  
 force = 2) +  
 scale\_x\_log10(labels = scales::comma\_format()) +  
 scale\_size\_continuous(range = c(3, 10)) +  
 scale\_color\_viridis\_c(option = "plasma") +  
 theme\_minimal() +  
 labs(  
 title = "CEO Compensation vs. Enterprise Value by HQ Location",  
 subtitle = "Bubble size indicates number of companies in each HQ location",  
 x = "Average Enterprise Value ($ Millions, log scale)",  
 y = "Average CEO Compensation ($ Millions)",  
 size = "Number of Companies",  
 color = "Avg CEO Comp ($M)"  
 )  
  
print(p\_location\_comparison)



The bubble size indicates the number of businesses in each place, and the bubble chart illustrates the correlation between average CEO remuneration and average enterprise value by region. Although enterprise value and CEO salary are highest in California and Illinois, the general trend is irregular indicating that enterprise value by itself is unable to adequately explain CEO compensation. The variance suggests that additional variables, such the nature of the industry and local economic circumstances, might have a big impact on pay.

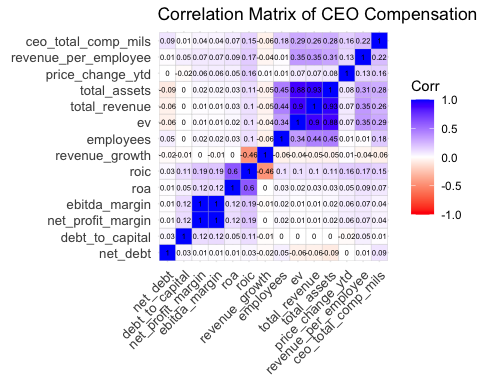
# Location boxplot to show distribution  
p\_location\_boxplot <- ggplot(imputed\_data %>%   
 filter(hq %in% location\_summary$hq),   
 aes(x = reorder(hq, ceo\_total\_comp\_mils, FUN = median),   
 y = ceo\_total\_comp\_mils,  
 fill = median(ceo\_total\_comp\_mils))) +  
 geom\_boxplot() +  
 coord\_flip() +  
 scale\_fill\_viridis\_c(option = "plasma") +  
 theme\_minimal() +  
 labs(  
 title = "Distribution of CEO Compensation by Location",  
 subtitle = "For locations with at least 5 companies",  
 x = NULL,  
 y = "CEO Compensation ($ Millions)",  
 fill = "Median CEO Comp ($M)"  
 ) +  
 theme(  
 legend.position = "none",  
 panel.grid.major.y = element\_blank(),  
 axis.text.y = element\_text(size = 9)  
 )  
  
print(p\_location\_boxplot)



For states with at least five businesses, the CEO compensation distribution by headquarters location is displayed in a box plot. It draws attention to the notable differences in CEO compensation between states, with Illinois and California showing more outliers and higher medians, indicating a greater range of salary. Several CEOs earn significantly more than the median, as evidenced by the existence of several outliers in several states.

### We will create a correlation matrix to check relationships between independent variables (avoid multicollinearity) and assess their correlation with our dependent variable (CEO compensation)

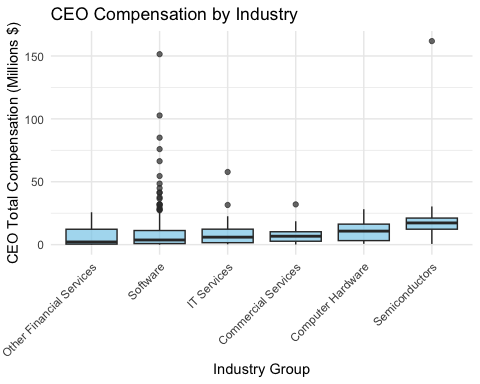
# Select only numeric columns for correlation analysis  
numeric\_vars <- ceo\_data[, sapply(ceo\_data, is.numeric)]   
  
# Compute the correlation matrix  
cor\_matrix <- cor(numeric\_vars, use = "pairwise.complete.obs")  
  
# Print the correlation matrix  
# print(cor\_matrix)  
  
# Optionally, visualize it with a heatmap  
library(ggcorrplot)  
ggcorrplot(cor\_matrix,   
 lab = TRUE,   
 lab\_size = 2, # Adjust text size  
 colors = c("red", "white", "blue"),   
 title = "Correlation Matrix of CEO Compensation Factors",  
 hc.order = TRUE, # Order by hierarchical clustering  
 tl.cex = 10, # Text label size  
 tl.srt = 45, # Rotate text labels  
 ggtheme = theme\_minimal()) # Use a cleaner theme



Overall, despite the research, we found that these variables actually don’t have the strongest correlation with CEO compensation. We will still explore these variables and see which within this we chose the following variables to focus on based on the above corr matrix: EV, revenue\_per\_employee, employees, roic, and price\_change\_ytd

### Visualizations

# Load necessary libraries  
library(tidyverse)  
library(ggplot2)  
library(car)  
library(stats)  
# Using the imputed\_data dataset from the provided information  
# 1. Box plots of CEO compensation by industry with ANOVA test  
# First, let's get the top industries by frequency to avoid too many categories  
top\_industries <- imputed\_data %>%  
 count(primary\_industry\_group, sort = TRUE) %>%  
 filter(n >= 10) %>% # filter for industries with at least 10 companies  
 pull(primary\_industry\_group)  
# Create a filtered dataset with only top industries  
industry\_filtered <- imputed\_data %>%  
 filter(primary\_industry\_group %in% top\_industries)  
# Create boxplot of CEO compensation by industry  
ggplot(industry\_filtered, aes(x = reorder(primary\_industry\_group, ceo\_total\_comp\_mils, FUN = median),   
 y = ceo\_total\_comp\_mils)) +  
 geom\_boxplot(fill = "skyblue", alpha = 0.7) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 labs(title = "CEO Compensation by Industry",  
 x = "Industry Group",  
 y = "CEO Total Compensation (Millions $)")



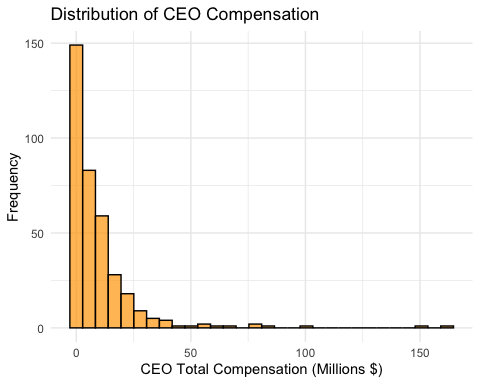
# Run ANOVA test to check if CEO compensation differs significantly between industries  
anova\_result <- aov(ceo\_total\_comp\_mils ~ primary\_industry\_group, data = industry\_filtered)  
summary(anova\_result)

## Df Sum Sq Mean Sq F value Pr(>F)   
## primary\_industry\_group 5 3900 779.9 2.745 0.0191 \*  
## Residuals 326 92607 284.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The CEO compensation distribution across several industrial groupings is displayed in the box plot. The greatest variance is seen in the software sector, where numerous outliers suggest that certain CEOs are paid disproportionately well. In contrast to other financial services, IT services, and commercial services, which have more closely clustered compensation values, the semiconductor and computer hardware businesses also exhibit comparatively higher median salary. This implies that CEO compensation is highly influenced by industry type, which may improve the regression model’s forecast accuracy.

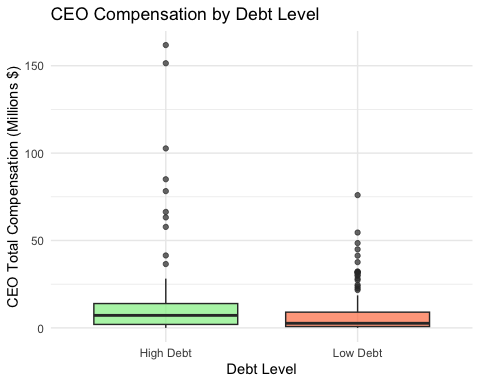
Regarding the ANOVA test, we have a p value of 0.0191 ( < .05), which simply confirms that there is a statistically significant difference in industry mean CEO compensations. We now know we can conduct meaningful analysis to understand how industry affects compensation.

# 2. Histogram showing distribution of ceo compensation   
ggplot(imputed\_data, aes(x = ceo\_total\_comp\_mils)) +  
 geom\_histogram(bins = 30, fill = "orange", color = "black", alpha = 0.7) +  
 theme\_minimal() +  
 labs(title = "Distribution of CEO Compensation",  
 x = "CEO Total Compensation (Millions $)",  
 y = "Frequency")



CEO salary is significantly skewed to the right, according to the histogram. A few extreme outliers earn over $100 million, whereas the majority of CEOs make between $0 and $10 million. This implies that although the majority of CEOs are paid moderately, the average is raised by a few high-paid CEOs. Given this skewness, median pay might be a better indicator of central tendency than mean.

# Understanding   
  
# Split companies into high and low debt using median as divider  
debt\_median <- median(imputed\_data$debt\_to\_capital, na.rm = TRUE)  
imputed\_data$debt\_level <- ifelse(imputed\_data$debt\_to\_capital > debt\_median, "High Debt", "Low Debt")  
# Visualize CEO compensation by debt level  
ggplot(imputed\_data, aes(x = debt\_level, y = ceo\_total\_comp\_mils)) +  
 geom\_boxplot(fill = c("lightgreen", "coral"), alpha = 0.7) +  
 theme\_minimal() +  
 labs(title = "CEO Compensation by Debt Level",  
 x = "Debt Level",  
 y = "CEO Total Compensation (Millions $)")



# Test for normality  
shapiro.test(imputed\_data$ceo\_total\_comp\_mils[1:5000]) # Limited to 5000 samples due to Shapiro-Wilk test limitations

##   
## Shapiro-Wilk normality test  
##   
## data: imputed\_data$ceo\_total\_comp\_mils[1:5000]  
## W = 0.52973, p-value < 2.2e-16

# Based on histogram and normality test, choose appropriate test  
# If skewed (likely), use Wilcoxon test - we choose this one  
wilcox\_result <- wilcox.test(ceo\_total\_comp\_mils ~ debt\_level, data = imputed\_data)  
print(wilcox\_result)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: ceo\_total\_comp\_mils by debt\_level  
## W = 21068, p-value = 3.035e-05  
## alternative hypothesis: true location shift is not equal to 0

# If normal distribution (unlikely), use t-test  
t\_test\_result <- t.test(ceo\_total\_comp\_mils ~ debt\_level, data = imputed\_data)  
print(t\_test\_result)

##   
## Welch Two Sample t-test  
##   
## data: ceo\_total\_comp\_mils by debt\_level  
## t = 2.841, df = 269.62, p-value = 0.00484  
## alternative hypothesis: true difference in means between group High Debt and group Low Debt is not equal to 0  
## 95 percent confidence interval:  
## 1.550813 8.552073  
## sample estimates:  
## mean in group High Debt mean in group Low Debt   
## 12.482111 7.430668

The findings of the Welch two-sample t-test and the Wilcoxon rank sum test show a significant difference in CEO compensation between companies with high and low debt levels. Given the extremely low p-values for both tests (3.03×10^5 and 0.00484), the null hypothesis is rejected and the statistical significance of the difference is confirmed. CEOs of high-debt companies typically make $12.48 million, but CEOs of low-debt companies make $7.43 million. This suggests that higher debt levels may be associated with higher CEO compensation, maybe as a result of increasing financial risk or the difficulty of managing high-debt companies.

According to the box plots, CEOs of businesses with large debt loads typically earn more money than those of businesses with smaller debt loads. High-debt corporations have a larger median salary, a wider compensation distribution, and more severe outliers. The results of the t-test, which indicated a statistically significant difference in remuneration between enterprises with high and low debt, are consistent with this. This implies that, maybe as a result of the greater complexity and risk associated with running such businesses, higher debt levels may be associated with higher CEO compensation.

Summary of Data & Methods

We used a variety of methods to examine CEO salary and find significant trends. To find underlying trends and aggregate related data points for exploration analysis, we employed clustering. While correlation matrices demonstrated the direction and strength of these interactions, histograms and scatter plots assisted in visualizing the compensation distribution and the link between various factors. We used the Wilcoxon Rank Sum test to analyze distributions, the Shapiro-Wilk test to examine data normality, and ANOVA to compare means across groups in order to determine statistical significance. In order to gain insight into how variables like sales, profit margins, and industry classification affect executive pay, we lastly employed Multiple Linear Regression to forecast CEO remuneration depending on firm success measures.

## Regression Analysis

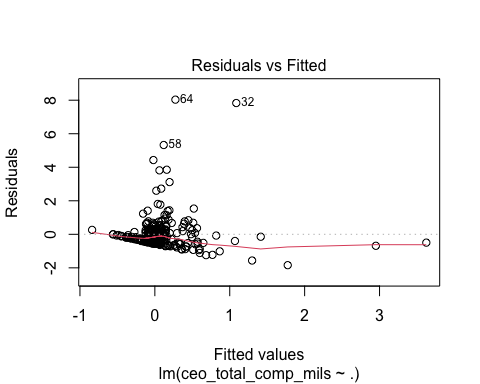
### Multiple Linear Regression: CEO Compensation ~ Financial Metrics

# Scaling Data to standardize

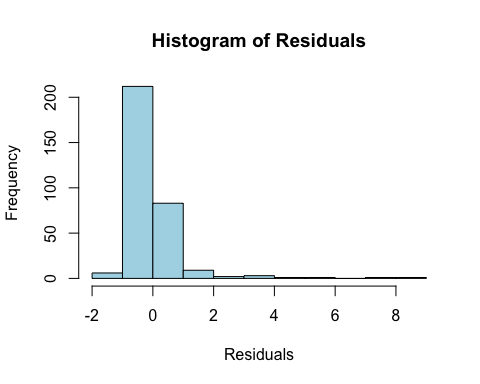
target\_vars <- c("ev", "roic", "revenue\_per\_employee", "employees", "price\_change\_ytd", "ceo\_total\_comp\_mils")  
  
ceo\_data\_scaled <- imputed\_data %>%  
 dplyr::select(all\_of(target\_vars))  
  
# Apply standardization  
ceo\_data\_scaled[target\_vars] <- scale(ceo\_data[target\_vars])  
  
lm\_full <- lm(ceo\_total\_comp\_mils ~ ., data = ceo\_data\_scaled)  
summary(lm\_full)

##   
## Call:  
## lm(formula = ceo\_total\_comp\_mils ~ ., data = ceo\_data\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8402 -0.4362 -0.1831 0.1319 8.0335   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.04954 0.05560 0.891 0.37368   
## ev 0.19243 0.06101 3.154 0.00177 \*\*  
## roic 0.09668 0.06973 1.386 0.16659   
## revenue\_per\_employee 0.15738 0.06451 2.440 0.01526 \*   
## employees 0.09660 0.05592 1.727 0.08507 .   
## price\_change\_ytd 0.10280 0.05640 1.823 0.06929 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9879 on 313 degrees of freedom  
## (48 observations deleted due to missingness)  
## Multiple R-squared: 0.1316, Adjusted R-squared: 0.1177   
## F-statistic: 9.483 on 5 and 313 DF, p-value: 1.959e-08

plot(lm\_full, which = 1) # Residuals vs Fitted



hist(residuals(lm\_full), main="Histogram of Residuals", xlab="Residuals", col="lightblue", border="black")



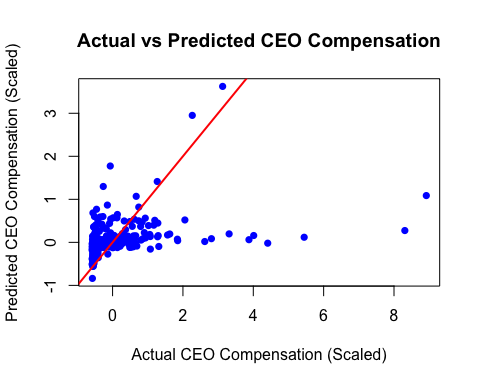
**What this full linear model tells us:** Multiple R-squared tells us that only 13.16% of the variance in CEO compensation is explained by the independent variables, which is extremely low and suggests that other important predictors are missing. The Adjusted R-squared is even lower, which tells us that the model does not generalize well.

The F-Statistic and low p-value tells us that the model is statistically significant overall, meaning that at least one predictor is likely useful.

**What the Residuals vs. Fitted Plot Tells Us** Ideally, the residuals should be randomly scattered around zero, but instead there is a cluster of points close to 0, and some extreme values. The outliers could be highly influencing the model and affecting generalization, and the non-linearity of the red line tells us that the model is probably not fully capturing all relationships.

**What the Histogram of Residuals Tells Us** Again, the ideal scenario for the residuals does not match the actual case. They should be normally distributed, but these are in a right-skewed distribution.

# Remove rows with NA in ceo\_total\_comp\_mils  
ceo\_data\_scaled <- ceo\_data\_scaled %>% drop\_na(ceo\_total\_comp\_mils)  
  
# Now create predicted values  
pred\_values <- predict(lm\_full, newdata = ceo\_data\_scaled)  
  
# Scatter plot  
plot(ceo\_data\_scaled$ceo\_total\_comp\_mils, pred\_values,   
 xlab = "Actual CEO Compensation (Scaled)",   
 ylab = "Predicted CEO Compensation (Scaled)",   
 main = "Actual vs Predicted CEO Compensation",  
 col = "blue", pch = 16)  
  
# Add a diagonal reference line  
abline(0, 1, col = "red", lwd = 2)



**Looking at the Actual Vs. Predicted Scatter Plot:**

If the model was perfectly trained, each blue dot would be on the red line (meaning the actual value = predicted value). If a blue dot is above the red line, then the model underestimated CEO compensation (actual > predicted). If a blue dot is below the red line, then the model overestimated CEO compensation (actual < predicted).

Looking at the scatter plot, there is more clustering at lower compensation values and more deviation from the line at higher compensation values. This tells us that the model struggles with accurately predicting higher-compensation CEOs.

Overall, we can see that linear modeling is likely not the best way to assess these relationships, and the next step in this project would be to train and assess a non-linear model.

# Results

Our initial analysis of CEO compensation by location showed that Illinois and California had the highest average pay, while Nevada had the lowest. However, headquarters’ location alone was not a strong predictor of CEO compensation, suggesting that market and financial conditions play a more significant role.

A correlation analysis identified five key financial factors that have the strongest relationships with CEO compensation:

* Enterprise Value (EV) – A measure of total company worth
* Revenue Per Employee – Efficiency metric linking revenue to workforce size
* Number of Employees – Indicator of company size
* Return on Invested Capital (ROIC) – Profitability measure
* Price % Change YTD – Stock market performance indicator

We then ran a multiple linear regression model using these factors. The model explained only 13.16% of the variance in CEO compensation (R² = 0.1316), indicating that many important predictors were missing. Among the variables, Enterprise Value and Revenue Per Employee had the strongest positive relationships with CEO pay. However, residual plots and skewed distribution of residuals suggested that a linear model might not be the best approach for this analysis.

A comparison of CEO pays by industry using an ANOVA test showed a statistically significant difference across sectors (p = 0.0191). CEOs in the software, semiconductor, and computer hardware industries tend to earn the highest median pay, likely to reflect the growth potential and competitive nature of these fields.

We also analyzed how CEO pay differs between high-debt and low-debt companies. On average, CEOs of highly leveraged firms earn significantly more ($12.48M) than those in low-debt firms ($7.43M) (p < 0.005). This suggests that financial complexity and risk management responsibilities may play a role in determining executive compensation.

A scatter plot comparing actual vs. predicted CEO compensation revealed that the model struggled to accurately predict higher compensation levels. The clustering of data points at lower compensation value indicates that our model may not fully capture the non-linear relationships affecting CEO pay.

# Discussion

Our findings highlight that while some financial factors—such as Enterprise Value and Revenue Per Employee—correlate with CEO pay, they don’t fully explain the variations. This suggests that other qualitative factors, such as board influence, negotiation power, and corporate governance policies, likely play a significant role.

The industry-level analysis reinforced that executive compensation varies widely across sectors, particularly in high-tech industries, where competition for top talent is intense. Additionally, our debt-level comparison suggests that CEOs managing high-debt firms may receive higher pay as compensation for taking on greater financial risk and complexity.

Given our insights gained, our study has some limitations:

1. Data Scope – Our dataset, sourced from PitchBook, focuses primarily on publicly available financial data and may not capture compensation structures in private companies.
2. Missing Variables – We did not include factors such as CEO tenure, stock ownership, or qualitative leadership attributes, which may be critical in explaining compensation differences.
3. Nonlinearity – The poor predictive performance of our linear model suggests that CEO pay is not linearly related to financial factors. More advanced techniques, such as decision trees, random forests, or machine learning models, could better capture complex relationships.

For investors, understanding the key drivers of CEO pay can improve executive pay evaluations and help assess whether compensation is justified relative to firm performance. For corporate boards, these insights can guide the development of more effective, performance-based pay structures. Policymakers and regulators can use this research to ensure fair pay practices and promote transparency in corporate governance.

Ultimately, our study underscores the complex and multifaceted nature of CEO compensation. While financial metrics matter, they only tell part of the story—other governance, industry, and negotiation-related factors likely play an equally important role.