**CEO vs. Worker Pay in Top 3000 US Companies**

   
   
 

**December 5, 2023**

Aakar Aryal, Amin Fara, Ashley Geary, Eugenia Galan, and Roshan Rai

MIT 5032

**Index:**

**Problem Statement**

**Hypotheses**

**Data Source and Description**

**Data Preprocessing Steps**

**Specific Approach/Analysis**

**Specific Outcomes/Insights**

**Summary and Conclusion**

**References**

**Problem Statement**

The CEO-worker pay ratio is a pivotal measure that sheds light on the internal compensation dynamics of companies and plays a crucial role in understanding the broader socio-economic landscape. The data correlates the median worker salary to the CEO salary and allows many relationships to be studied. This study focuses on exploring the intricacies of CEO-worker pay ratios within the top 3000 companies in the United States, considering diverse factors such as gender, industry affiliation, and corporate rankings.

The data set aims to explore and visualize the relationship between CEO pay and worker pay, compare different industries and companies, and uncover patterns and trends in the data. This dataset is an important resource for anyone interested in promoting pay equality and advocating for fair wages.

The industries under consideration are: "Health Care," "Materials," "Industrials," "Consumer Discretionary," "Information Technology," "Real Estate," "Financials," "Consumer Staples," "Communication Services," "Utilities," and "Energy”. Consumer discretionary sector are companies that produce non-essential goods and services like Ford, Netflix and Tesla. Consumer staples are companies that sell products that can be considered essential, for example, Walmart, and Proctor&Gamble.

**Hypothesis**

**Hypothesis #1**

Male CEOs in the Information Technology have a higher pay ratio than Female CEOs in the Information Technology. They are inversely correlated.

Analyzed using Categorical Nominal: “M” and “F”

**Hypothesis #2**

Financial companies with a higher S&P Rank, there will be a higher CEO-worker salary ratio. There is a positive correlation relationship.

Analyzed using categorical ordinal data, where S&P Rank or R 3000 ranking is highest at the top (#1) and lowest at 3000.

**R-value: -0.05953562 (Negative corelation)**

**Hypothesis #3**

Healthcare companies have the highest median worker pay, and the lowest CEO-worker salary ratio. They are inversely correlated. – Ashley

Analyzed using continuous Interval

R-Value: -0.3608818 (Strong negative correlation)

**Hypothesis #4**

Female CEOs in the Top 100 companies have higher CEO-worker salary ratio than the Male CEO’s in the top 100. Companies with Male CEO’s have a 3:1 ratio and have a higher median work pay. – Amin

**Continuous Discrete (2 plots with F v M)**

**R-Value: 0.0008506226 (no correlation)**

**Hypothesis #5**

Industrial companies with a higher pay ratio indicate in general a higher CEO-worker salary. This shows a positive correlation relationship. Additionally, CEOS in the industrial sector are heavily male with only 6% of CEOS represented in the dataset being female. – Aakar

Continuous Interval

R-value: 0.4349431 (strong positive correlation)

**Data Source and Description**

The data source was a Kaggle dataset for CEO vs Worker Pay in Top 3000 US Companies. This dataset included salaries from Russell 3000 and S&P 500 companies. It is a public data set available for use.

Using functions like “dim” and “colnames” the data dimensions and column names can be defined. The data had 3000 companies and 8 columns, but after dropping some rows the data ended up with 2,175 unique values. The eight columns were the index, ticker, company name, median worker pay, pay ratio, CEO name, salary, and industry. This is further described in the preprocessing steps.

The limitations from this data set are that there may lack of data completeness, resulting in NULL values. Additionally, while the data was collected in 2023, there is no indication that the median pay for each worker is the latest average, this could result in inaccurate data quality. There also may be some biases, from the selection process of the companies into the top 3000 companies by Russell. There may be a bias towards companies that were nominated in the selection process.

**Data Preprocessing Steps**

The preprocessing of the data from the dataset was done methodically. The first step was seeing the parameters of the data. Starting with the “summary” function to determine the dataset. This was a quick way to identify if there were any NULL values in the data. There were 28 NULL values in some of the rows, so those rows were dropped. This step was helpful in seeing the max and min salaries of the CEO and the median worker pay. The median worker pay is interval continuous.

The second step was cleaning the data, the data had 28 NULL values for the salaries column. The team decided that the NULL values should be discarded, since the values would not show up in the analysis. This resulted in 2175 unique values for extrapolation.

The next step was the normalization of the data. For the salary, it was in the currency format for USD. The “$” sign and other symbols on the salary was removed and converted to an integer using “as.numeric” and “gsub” function to ensure that the data can be used for analysis. In order to use the data properly, the data was converted to integer, and additional columns were added with “\_int” added to the column name to signify it has been converted to a whole number. This was done to be consistent with the data.

Additionally, the pay ratio had to be converted to a whole number. Since the dataset was Number:1, the last two characters “:1” were removed, and the assumption was to add that it was a ratio to one. This normalized the pay ratio column.

A few data columns were added, including company size, ESG risk levels and categorization. This would have added to the richness of the data, since ESG has been of high interest in recent years. However, due to numerous null values, the columns could not be used.

**Specific Approach/Analysis**

Before proceeding with the specific approach, we took a big-picture snapshot of the data and extrapolated the average and max CEO salary and average and max worker median pay. Additionally, a pie chart was created to ensure that all the top 100 companies – with the highest CEO-worker pay ratio were identified. The highest count was in the “Consumer Discretionary” industry and “Health Care” had the highest count in the lowest 100 companies, indicating that the “Consumer Discretionary” had the highest pay ratio between CEO and workers, with “Health Care” having the lowest pay ratio.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **S&P 500 ($)** | **R3000 ($)** | **Combined ($)** |
| **Average CEO Salary** | 18,582,399 | 8,316,591 | 10,639,856 |
| **Max CEO Salary** | 296,247,749 | 834,959,367 | 834,959,367 |
| **Average Worker Median Pay** | 84,866 | 95,180 | 92,846 |
| **Max Worker Median Pay** | 295,884 | 753,041 | 753,041 |

A graph of a person and person

Description automatically generated

**Proving Hypothesis 1:**

To determine the relationship between CEO-worker Salary Ratio in the Information Technology (IT) industry, the plot was graphed on R using a violin plot. Since the data was categorical nominal data, this plot was chosen to study the relationship. The relationship found was that Male CEOs do have the highest CEO-worker pay ratio, but the majority of Male CEOs and Female CEOs have an average ratio of 234:1. Interestingly, for female CEOs, there are two additional groups, both of which have a higher ratio, approximately 600:1 and 1800:1, Therefore, it can be concluded that the hypothesis is true that male CEOs in the Information Technology sector have a higher pay ratio compared to female CEOs. The correlation can be analyzed using Cramer’s V, but since it is a categorical nominal value, the violin plot shows the relationship with the frequency.

A graph of a graph showing a number of blue dots

Description automatically generated with medium confidence

**Proving Hypothesis 2:**

A graph of a scatter plot

Description automatically generated

This scatterplot visually represents the relationship between CEO-to-Worker Pay Ratio and Median Worker Pay in a dataset. On the x-axis, we have the CEO-to-Worker Pay Ratio, y-axis displays Median Worker Pay, an interval continuous variable representing the annual median pay for employees. There is a color representation for the rank as it is categorical ordinal data. The correlation coefficient, calculated as -0.08174, suggests a very weak negative correlation between CEO-to-Worker Pay Ratio and Median Worker Pay. In other words, as the CEO-to-Worker Pay Ratio increases or decreases, there is a slight tendency for Median Worker Pay to move in the opposite direction, but the relationship is minimal. It's important to note that correlation does not imply causation, and various factors may influence these variables independently. The negative correlation coefficient indicates a subtle inverse relationship, but the strength of the association is negligible.

|  |  |
| --- | --- |
| A graph of a graph with blue dots  Description automatically generated with medium confidence | A graph with many dots  Description automatically generated with medium confidence |

Moreover, our initial hypothesis stated that there would be a positive correlation between the Russell 3000 ranks of financial companies and their CEO-worker salary ratios. Upon examination of the visual representation, it is evident that the darker dots, indicative of higher ranks, tend to exhibit a slightly elevated pay ratio, thereby supporting our assumption. Nevertheless, due to some ambiguity in the image, we sought additional confirmation through another graph. The resulting correlation coefficient of -0.05698869 reveals a weak negative correlation. This implies that as the Russell 3000 rank increases, the CEO-worker salary ratio experiences a slight decrease. Put simply, as the rank progresses from 1 to 3000, the pay ratio diminishes. In other words, we can express this correlation as follows: "as the rank decreases, the pay ratio also decreases," aligning with our original hypothesis.

However, it's crucial to consider the null hypothesis (H0) in our analysis. The null hypothesis might posit that there is no significant correlation between Russell 3000 ranks and CEO-worker salary ratios. The weak negative correlation we observed could be within the range of what would be expected by random chance or other factors not considered in our hypothesis. In statistical terms, our observed correlation might be statistically indistinguishable from zero. While our initial observation seems to suggest a negative trend, it's essential to acknowledge the possibility that this observed correlation could have occurred by chance.

This scatter plot was created to determine the relationship between S&P Rank and CEO-Worker Pay Ratio indicates that CEO-Worker Pay ratio falls under 250:1 for most instances. This data is interval discrete.

A graph of a worker salary rate

Description automatically generated

A graph of a company

Description automatically generated

A graph of a salary rate

Description automatically generated with medium confidence

**Proving Hypothesis 3:**

The hypothesis here is proven to show that the data is rightly skewed, giving us more clarification regarding the salary distribution in the healthcare sector. In our graph, while the salary gap is smaller, there exists a larger concentration of healthcare companies with lower CEO-worker salary ratios compared to the median. Within our dataset, healthcare companies are distinct from other industries in offering higher median worker pay while maintaining a notably lower CEO-worker salary ratio. Thus, on average, employees working for healthcare companies receive relatively higher wages compared to workers in other industries in our dataset. Further, the compensation gap between CEO and worker pay is smaller. The observed negative correlation reported of -0.36 between median worker pay and the CEO-worker salary ratio supports this relationship. Moreover, as median worker pay is typically higher in healthcare companies, the CEO-worker salary ratio is lower, validating the inverse relationship. Our data and graphical representation show healthcare companies prioritize providing higher wages to their employees while minimizing the compensation gap between employees and CEOs when compared to the industries studied in our dataset. This key takeaway highlights the uniqueness of the healthcare industry in its compensation practices compared to other sectors.

A graph with blue and red dots

Description automatically generated

**Proving Hypothesis 4:**

Female CEOs in the Top 100 companies have higher CEO-worker salary ratio than the Male CEO’s in the top 100. Companies with Male CEO’s have a 3:1 ratio and have a higher median work pay.

The investigation sought to substantiate the hypothesis that female CEOs in the Top 100 companies would exhibit a higher CEO-worker salary ratio than their male counterparts. Through a meticulous analysis of salary data and the construction of dual-axis plots, our study aimed to uncover distinct patterns within each gender category. Contrary to the initial assumption, the graphical representation revealed that companies with male CEOs consistently maintained a 3:1 CEO-worker salary ratio, and the median worker pay was higher compared to those led by female CEOs. The statistical analysis, encapsulated by an R-Value of 0.0008506226 denoting no correlation, dispels the notion that CEO gender serves as a significant predictor for variations in salary structures. In essence, our findings challenge the presumed link between CEO gender and salary ratios, underscoring the need for a nuanced understanding of the multifaceted factors influencing corporate compensation practices.

A graph with many dots

Description automatically generated

**Proving Hypothesis 5:**

Industrial companies with a higher pay ratio indicate in general a higher CEO-worker salary. This shows a positive correlation relationship. Additionally, CEOS in the industrial sector are heavily male with only 6% of CEOS represented in the dataset being female.

A scatter plot was created to show the relationship of Pay Ratio to Median Worker Salary in the Industrial sector of our data. The pay ratio falls under 250 for most of the data with a positive correlation being present and the data interval continuous. For the industrial sector, we saw that tendency present for as the pay ratio increases, there being a tendency for median worker salary to also increase. The conclusion can be made based on that that companies with higher pay ratios in the industrial sector tend to have higher median worker salaries. The industrial sector among the top companies we found in our dataset is severely lacking female representation due to there only being a handful of CEOs who are female. The conclusion can be made that diversity and inclusion in the industrial sector is greatly lacking. The data shows the R-value of 0.4349431 which is a strong positive correlation indicating that the hypothesis rings generally true.

**Specific Outcomes/Insights**

On average, among the Russell 3000 companies, 1 CEO Salary = 10 Workers Median Pay

The highest CEO salary is in the Communication Services industry with $834,959,367, The Trade Desk, Inc. Interestingly, the lowest median wage paid companies are companies like Scholastic, and Liberty Latin.

According the bizjournals, there are 459 male CEOs and only 42 female CEOs in the S&P 500 companies. However, a recent report proves that female CEOs outperform their male counterparts, especially during COVID. And females generally have a lower CEO-worker pay ratio on average as proven in hypothesis 1.

|  |  |
| --- | --- |
| **Top 100 Companies by Industry** A pie chart with different colored circles  Description automatically generated | **Bottom 100 Companies by Industry**  A colorful pie chart with text  Description automatically generated |

**Top 2 largest gap:**

Consumer Discretionary, 490.4:1

Communication Services: 297.2:1

The companies above represent the least number of instances, by count, where there exists a high CEO-worker pay ratio.

**Bottom 2 most equal gap:**

Energy, 80.5:1

Utilities, 78.4:1

The companies above represent the least number of instances, by count, where there exists a low CEO-worker pay ratio. These findings indicate stable and consistent compensation practices based on industry-specific business models.

IX. Specific Outcomes and Key Learnings   
● Despite the authors efforts to identify and develop a project with manageable scope and established statistical framework, it was quickly apparent that the health inspection data was very complex and require extensive preprocessing. While validation of the dataset did occur, earlier interaction with the   
data would have been helpful.   
● Only one business in the data frame of 3061 businesses is risk category 1. It is a student store at   
Bellevue High School. Because of this, the findings in this analysis will not be applicable to risk   
category 1 businesses.   
● 339 (95.2%) businesses in risk category 2 are school lunch programs. These inspection records do not   
have grades assigned to them. Further, the high prevalence of this one, non-commercial business type   
may be skewing the curve for other risk category 1 and 2 businesses in these zip codes. If category 2   
businesses are of significant interest to this study, additional questions should be asked to confirm if   
they are included in the risk category 1 and 2 zip code curves.   
● The difference in scores between the “Downtown” and “University” districts may be in part related to   
the presence of many fraternity houses in zip code 98105. These are considered businesses by the   
health department and were a category specifically commented on by Chris Skilton during the phone   
interview. 6 Additional research is warranted to determine the effect of this business category.   
● In all Seattle, Shoreline, and Bothell, the types and frequency of violations were fairly consistent with   
each other and the overall sample. This may be due in part to the large influence of Seattle-based   
businesses represented in the dataset.

**Specific Outcomes**

**Hypothesis:**

* Male CEOs in the IT industry tend to have a higher CEO-worker pay ratio compared to their female counterparts.
* There is a positive correlation between S&P Rank and CEO-worker pay ratios.
* The healthcare sector prioritizes higher median worker pay with a smaller CEO-worker pay gap.
* There is no significant correlation between gender and CEO-worker pay ratios within the top 100 companies.
* There is a positive correlation between pay ratios and CEO-worker salaries in the industrial sector.
* Data processing and manipulating was huge part of the project. There were other sources than Kaggel which were used to check the hypothesis.

**External sources + data conclusion**

* **Gender gap**
* **Industry discrepancies**

The data revealed a persistent gender pay gap among CEOs in the industries analyzed. The data showed outliers indicating progress among female CEOs, however, further research revealed this gap widened during the pandemic. Female C-suite pay fell to a record-breaking low as a percentage of pay earned by male CEOs that has not been seen since 2012. From 2012 to 2020, male CEOs gained a 27% increase in pay compared to female CEOs, who achieved a 10% increase. By 2020, female CEOs earned 75% of male CEO pay, which is a decrease compared to 88% recorded in 2018 (4). These insights are derived from proxy statements filed by the S&P500 companies, also revealing that the pay discrepancy could be due to the stock-related awards that men out-earn at a significant margin compared to women. The pay gap could also stem from early careers, where women graduating with bachelor’s degrees typically earn less than men, where they experience biases that influence compensation across varying fields. Additionally, certain career paths are viewed as more appropriate and effective to climbing to the CEO level. Pay transparency has helped to narrow the pay gap within companies for individual employees and managers, but female CEOs are still facing bias. Today, women hold over half of management-level positions and account for half of the U.S. population, however, are underrepresented in CEO roles. Men are outnumbering women in these executive-level leadership positions.

We identified top industries where employees earn the highest pay, resulting in lower CEO-to-worker pay ratios. The CEO-worker pay distribution in the Health Care industry can be explained by the company size or by the presence of employees pursuing more specialized roles with advanced degrees. For instance, employees working for research companies in the healthcare industry, such as the top-worker-paid Kodiak Sciences, earn top salaries, with the highest salaries earned in California, New Jersey, and Massachusetts. These high salaries further decrease the CEO-worker pay ratio.

**Key Learning**

Throughout this project we explored the dataset relation to CEO-worker pay ratio. This data provided insightful observation especially regarding the gender pay gap among CEOs in various industries. We explored many different industries to get the most variety of data. Although, as mentioned earlier, the data only comes from the top 3000 companies from Russel while there are many more companies across the U.S. which may give a different picture. Despite this, we tried to get the most out of the data we had and made some interesting findings.

**References:**

1. <https://www.kaggle.com/datasets/salimwid/latest-top-3000-companies-ceo-salary-202223>
2. <https://www.barrons.com/market-data/indexes/rua>
3. <https://www.bizjournals.com/bizwomen/news/latest-news/2023/04/you-are-more-likely-to-have-a-ceo-named-david-than.html?page=all#:~:text=According%20to%20a%20recent%20report,from%20female%2Dled%20companies%20vs>.
4. [Gender Pay Gap for Executives Grew During Pandemic (shrm.org)](https://www.shrm.org/resourcesandtools/hr-topics/compensation/pages/gender-pay-gap-for-executives-grew-during-pandemic.aspx)

Link to project folder: <https://teams.microsoft.com/_#/school/FileBrowserTabApp/General?threadId=19:2ukOcOVueJlahpyaRaOs6fzg0-vyuGMHqjR4JbwqiN81@thread.tacv2&ctx=channel>

Link to data: <https://sooners.sharepoint.com/:x:/s/AnalyticsprogR/ERWwVt1YKa9Hu-0TC88Gg7YBSiJ_xvg0evxSYwfw5youOA?e=ZYXJa5>

Link for S&P 500 ranks <https://www.slickcharts.com/sp500>

**Appendix:**

Top 100:

|  |  |
| --- | --- |
| **Industry** | **Count** |
| Consumer Discretionary | 54 |
| Information Technology | 18 |
| Consumer Staples | 8 |
| Communication Services | 7 |
| Health Care | 4 |
| Industrials | 4 |
| Materials | 3 |
| Financials | 2 |

Bottom 100:

|  |  |
| --- | --- |
| **Industry** | **Count** |
| Health Care | 29 |
| Financials | 18 |
| Information Technology | 17 |
| Real Estate | 12 |
| Industrials | 6 |
| Consumer Discretionary | 5 |
| Communication Services | 5 |
| Energy | 4 |
| Materials | 2 |
| Consumer Staples | 1 |
| Utilities | 1 |