MGT 6203 Group Project

Progress Report

A Statistical Analysis of the Gender Pay Gap

GitHub: MGT-6203-Fall-2023-Canvas/Team-26

Team 26

Marissa Robinson

Jamie Hernandez Kluesner

William D'Onofrio

Sanjay Lindsay

Emmett Drake

Background

This year marks the 103-year anniversary of the ratification of the 19th Amendment. This amendment paved the way for a major change in women's rights and roles in the United States. However, the journey to full gender equality is not complete. Despite the passing of the 19th amendment, the women's rights movement of the 1960's, and increased attendance of women in college, there is still a discrepancy between what a man and woman get paid for the same job. The primary objective of this analysis is to unravel the factors most closely linked to this pay disparity. From a business perspective, addressing the gap isn't merely a moral imperative. Companies striving for excellence in the competitive market must not only offer appealing financial packages but also champion true equality. Ensuring equal pay for equal work not only fosters a diverse workforce but also attracts top tier talent, positioning businesses optimally for sustained success. Competitive markets must not only offer appealing financial packages but also champion true equality. Ensuring equal pay for equal work not only fosters a diverse workforce but also attracts top tier talent, positioning businesses optimally for sustained success.

Initial hypothesis

In our investigation into the nuances of the gender pay gap, several hypotheses guide our investigation. First, we hypothesize that factors such as job performance evaluation scores, education level, and seniority (years of experience) will be highly correlated with the pay disparity between men and women. Specifically, we expect that women with lower performance scores, lower education levels, and fewer years of experience will experience a greater pay disparity compared to men. Turning our attention to departmental and role-specific disparities, we hypothesize certain departments and job titles—potentially male-dominated ones like engineering or management—to exhibit heightened gender pay gaps relative to roles where gender representation is more balanced. Additionally, we hypothesize that advanced education acts as a buffer against the gender pay gap. That is, women equipped with advanced degrees, such as master's or PhDs, might navigate a lesser pay gap in contrast to those with just high school diplomas or bachelor's degrees. Lastly, in examining age-related variances, our hypothesis leans toward older age brackets, perhaps those aged 40 and above, as potentially harboring the most substantial pay gaps. This is predicated on the notion that enduring historical gender biases have left their mark on these cohorts, while younger age groups, molded by evolving societal norms, may present a narrowing gap.

The Data

Data Preprocessing & Exploration:

Initially, we read the dataset "Glassdoor Gender Pay Gap.csv", which contains information such as Job Title, Gender, Age, Performance Evaluation, Education, Department, Seniority, Base Pay, and Bonus. From the summary statistics, we observed that our dataset has:

- Ages ranging from 18 to 65, with a median age of 41.
- Performance evaluations scores between 1 to 5, averaging around 3.
- Base salaries ranging from 34,208 to 179,726 with a median base salary of 93,328.
- Bonuses ranging from 1,703 to 11,293, with a median bonus of 6,507.
- Furthermore, we created a Salary variable by summing the BasePay and Bonus for each record. Upon inspecting for missing values, our analysis confirmed the absence of any missing data in our dataset.

Exploratory Data Analysis (EDA):

Our preliminary data analysis suggested that the salary distribution might differ based on gender. In figure 1, we present a graphical representation of the salary distribution grouped by gender.

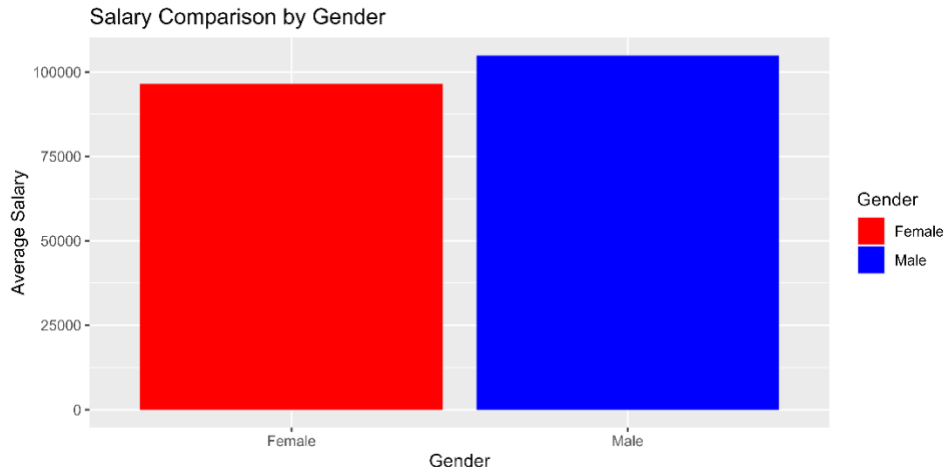

Figure 1: Comparison of Salary distribution by Gender

Outlier Identification & Analysis

Using the Interquartile Range (IQR) method, three potential salary outliers were identified. All these outliers were males working in managerial roles or in IT. Interestingly, upon checking if there are any female records with similar attributes as these male outliers, we found none. This observation intensifies the importance of our analysis, particularly in the context of exploring the potential gender pay gap. Given the significance of these findings, we have chosen to retain these outliers for the subsequent phases of our research.

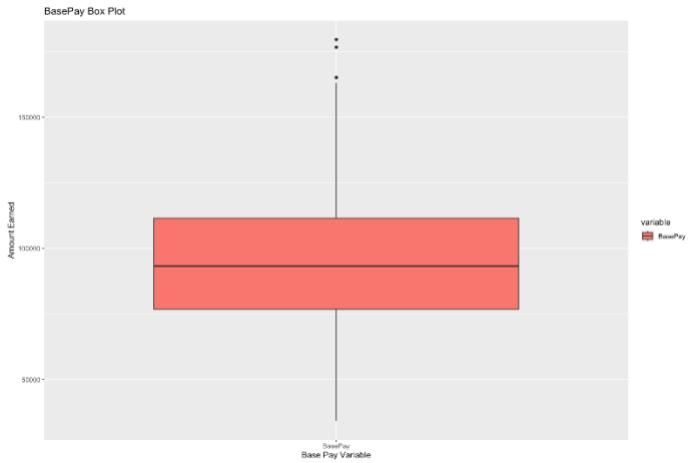

Figure 2: Boxplot summarizing the outlier analysis

Income Differentials: A Linear Regression Approach

Regression Analysis for Predictor Significance

Using simple linear regression models, we evaluated several independent variables: Age, Job Title, Education, Department, Seniority, Gender, and Performance Evaluation. The significant predictors at the 90% confidence level for Salary in a single predictor model turned out to be Age, Education, Department, Seniority, and Gender.

Simple Linear Regression (SLR) Analysis for Gender

Our SLR analysis with Gender as the predictor showed that, on average, males earn $8,502.00 more than females. This difference was statistically significant at the 99% confidence level, and the model's F-test of 29.24 (p-value 8e-08) indicates strong evidence to reject the null hypothesis. Residual plots from this model suggested that the assumptions of normality and constant variance are reasonably met. Using Cook's distance, we did not identify any overly influential points in our model, suggesting that our dataset is robust for the analysis.

Our analysis highlights a potential gender pay gap within the dataset. Males, on average, earn approximately 8.8178% more than females. This difference is statistically significant and warrants further exploration of this disparity. Our other predictors may provide additional detail. We will proceed with Logistic Regressions for further analysis.

Income Differentials: A Logistic Regression Approach

In our analysis, we utilized logistic regression to determine the factors influencing whether an employee's salary exceeds the median value. The results indicate several notable findings:

- Age emerged as a significant predictor ($p < 2e-16$), with older employees being more likely to earn above the median salary. Specifically, with each year increase in age, the log odds of having a high salary increase by 0.18376.
- Among the various Job Titles, Managers ($p < 1.45e-15$) and Software Engineers ($p < 5.85e-05$) demonstrated a particularly higher likelihood of earning above the median salary. Conversely, Marketing Associates were significantly less likely ($p < 3.09e-11$) to earn more than the median. Other titles like Driver, Graphic Designer, IT, Sales Associate, and Warehouse Associate did not exhibit a robust association with the high salary outcome.
- Employees with Masters ($p = 0.000444$) and PhD ($p = 0.001714$) degrees were more likely to earn above the median, compared to those with other education levels.
- In terms of Department, individuals in the Engineering ($p = 0.003305$) and Sales ($p = 0.000920$) departments exhibited a higher propensity for higher earnings, while the Operations department did not show any significant impact.
- Seniority played a crucial role ($p < 2e-16$) in determining salaries. For each unit increase in seniority, the log odds of having a high salary increased by 1.93493, indicating that more senior employees are more likely to earn above the median.
- Performance Evaluation (PerfEval) scores were also positively associated ($p = 0.018993$) with earning above the median, indicating that employees with better evaluations tend to have higher salaries.
- Gender, in this analysis, did not demonstrate a significant association with the likelihood of earning above the median salary.

To evaluate the model's predictive power, we assessed its performance on the training data. The confusion matrix revealed that our model correctly classified 432 individuals earning below the median and 438 earning above. However, it misclassified 62 individuals as earning below the median when they were earning above, and 68 individuals as earning above the median when they were in fact below. These results indicate a promising model fit with areas of potential refinement.

Decoding the Gender Pay Gap: A Tale of Two Regressions

In examining the intricacies of the gender pay gap, our analysis utilized two distinct yet complementary statistical approaches: Simple Linear Regression (SLR) and Logistic Regression (LR). The SLR, focusing on absolute salary values, revealed a discernible gender disparity in wages. This suggests that, on average, and after adjusting for other factors, there exists a measurable wage disparity between genders. In contrast, the LR, which pivots on the median salary as a benchmark, indicated that gender does not significantly predict the likelihood of an individual earning above or below this median value. This nuanced finding emphasizes that while there may be an overall wage difference between genders, this disparity does not necessarily manifest around the

median salary range. Collectively, these results underscore the complexity of the gender pay gap issue. It is not solely about broad disparities but also about understanding where in the salary distribution these disparities are most pronounced. By leveraging both analytical techniques, we gain a holistic perspective, recognizing that the gender pay gap can simultaneously be evident in average wages while being less discernible around median salary values.

Exploring Gender Interaction Effects

In an additional model, interaction terms were incorporated between gender and each respective predictor. The significance of such interaction terms indicates differential effects of one predictor on the outcome based on the levels of another predictor.

Two notable interactions emerged: Department (Engineering) x Gender (Male) and Department (Management) x Gender (Male). Both interactions were statistically significant, indicating variations in the gender pay gap across these specific departments. In essence, male affiliation in the Engineering and Management departments was associated with a heightened likelihood of earning a higher salary in comparison to the designated reference group.

Further examination revealed notable interactions with the Job Title (IT) x Gender (Male) and Job Title (Sales Associate) x Gender (Male). While the interaction involving the IT designation approached conventional significance threshold ($p = 0.05151$), the interaction with the Sales Associate designation was statistically significant. This illuminates potential heterogeneity in the gender pay gap among these job roles.

Conversely, some interactions, exemplified by Education x Gender (Male), were not statistically significant, indicating that within the dataset, the magnitude of the gender pay gap did not exhibit considerable variability across educational attainment levels.

Works Cited

1. 19th Amendment to the U.S. Constitution: Women's Right To Vote (1920), National Archives and Records Administration. Available at: https://www.archives.gov/milestone-documents/19th-amendment
2. Wolfers, J. (2021, November 23). *More women than men are going to college. that may change the economy.* The New York Times. https://www.nytimes.com/2021/11/23/business/dealbook/women-college-economy.html
3. Second Wave Feminism Primary Sources & History | Gale. (n.d.). https://www.gale.com/primary-sources/womens-studies/collections/second-wave-feminism#:~:text=The%20second%20wave%20feminism%20movement,spread%20to%20other%20Western%20countries

Next Steps

- Deeper dive into interactions

    - Further explore the significant interaction terms to better understand their implications
- Gender Breakdown by Role and Department
    - Given that gender interacts differently across departments and roles it may be useful to look at them more closely, which can also help explain some interaction effects
- Logistic Regression needs model validation and possible tuning
- Present potential Solutions and Recommendations
- Create more useful data visualizations