SFU SIMON FRASER UNIVERSITY

Simon Fraser University

STAT403 - Intermediate Sampling and Experimental Design

# Exploring the Gender Pay Gap

Author: Firas Fakih

April 17th 2021

# Table of Contents

# Abstract

To investigate the gender pay gap, data consisting of salaries taken from Glassdoor and analyzed using stacked bar plots, box plots, waffle charts. A regression model was built to predict the future trend of salaries for males and females and how they relate to each other. It was concluded that there was a significant improvement in reducing the wage gap, and not only do females make almost as much as males, they even exceed their pay in certain fields. In jobs such as Software Engineering, we can see a significant gap between males and Females, where males make almost $12,000 more than what Females make. The regression model also highlights the fact that male and female pay exhibit a strong positive correlation.

# The Problem

The work industry has grown rapidly over the past few years with the advancement of Data Science, Cyber Security, Big Data, etc. But companies still can't seem to avoid gender pay inequalities (Lindstaedt, 2019). This study is being carried out to investigate whether or not modernity has forced companies to reduce the wage gap, more specifically the question is whether the gender pay gap exists within different Age groups, Seniority levels, and Job Titles.

**Objective: Does the Gender Pay Gap still exist in our world today?**

# The Data

The dataset has been taken from Glassdoor and focuses on income for various job titles based on gender. This data set will be helpful in identifying the depth of the gender-based pay gap.

The data set has the following features:

- **Job Title, Gender and Age**
- **Performance Evaluation:** How the Employee is performing in the company. Ranking from 1 to 5, 5 being the best evaluation
- **Education:** The highest level of education attained by the employee
- **Department**
- **Seniority:** The level of seniority the employee has at the company, ranking from 1 to 5, 5 being the highest.
- **Base Pay and Bonus Pay**

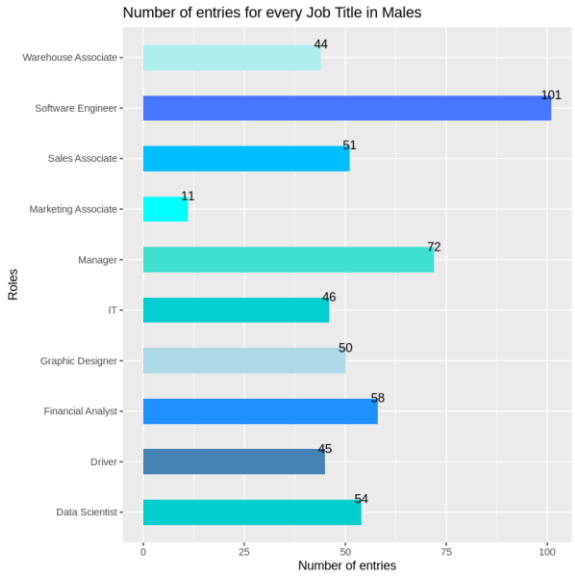
# Preprocessing and Exploratory Data Analysis

To help understand the data better, ggplot and dplyr libraries were used. ggplot2 is a plotting package that makes it simple to create complex plots from data in a data frame. It provides a more programmatic interface for specifying what variables to plot, how they are displayed, and general visual properties. ("*Data visualization with ggplot2").* dplyr is a new package which provides a set of tools for efficiently manipulating datasets in R. dplyr is the next iteration of plyr, focusing on only data frames. dplyr is faster, has a more consistent API and is easier to use. (*Introducing Dplyr*, 2014)

## Grouping Data by Job Title for males and females

Out of 1000 entries, the number of employees based on Job Title are fairly equal across all jobs except for Software Engineer (101 Male / 8 Female, Managers (72 Male / 18 Female) as well as Marketing Associates (11 Male / 107 Female), as shown in Figures 1 and 2 below.

*Figure 1: Number of entries by Job Title for Males*



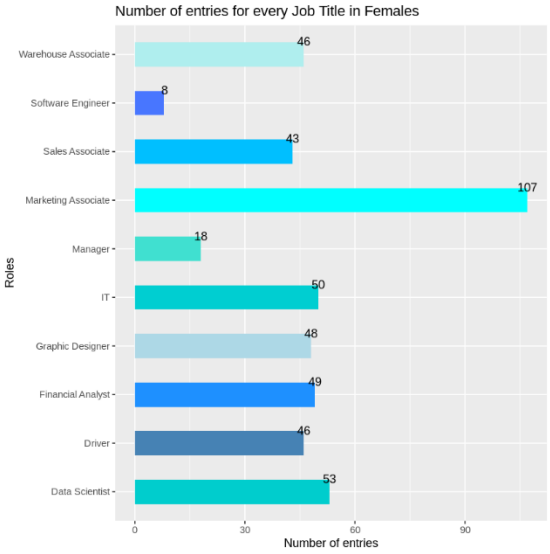




*Figure 2: Number of entries by Job Title for Females*

## Education and Seniority Distribution

After plotting 'waffle' charts shown in Figures 3, 4, 5 and 6 below, we observe that the data is well distributed and captures roughly the equal number of female and male entries based on Education and Seniority. Males see a slight increase in graduate studies shown in Figures 3 and 4 compared to females, as well as a slightly higher number of superior employees, shown in Figures 4 and 5.

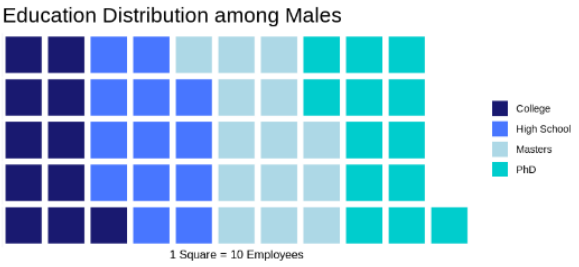



*Figure 3: Waffle chart showing education Distribution for males*

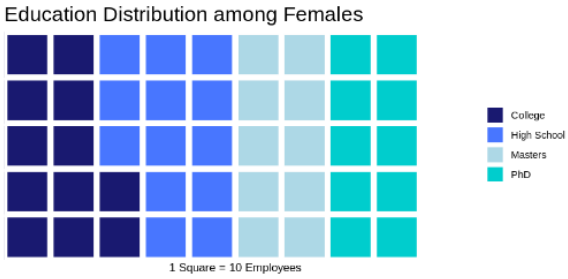



*Figure 4: Waffle chart showing education Distribution for females*

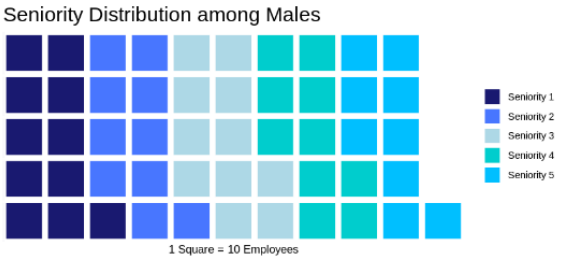



*Figure 5: Waffle chart showing Seniority Distribution for males*

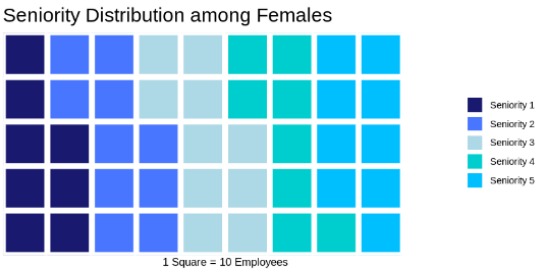



*Figure 6: Waffle chart showing Seniority Distribution among females*

# Statistical Analysis

## Comparing Means

We first start our Analysis by comparing the means of the Base Pay and the means of the Total Compensation (Base Pay + Total Pay) between Males and Females. Figure 7 clearly shows that on average, males earn a higher base salary than Females, earning $8,515 more. After adding bonus pay, Figure 8 shows that on average males still earn more than Females, earning $8,502, indicating that Females earn an insufficiently higher bonus pay compared to males.

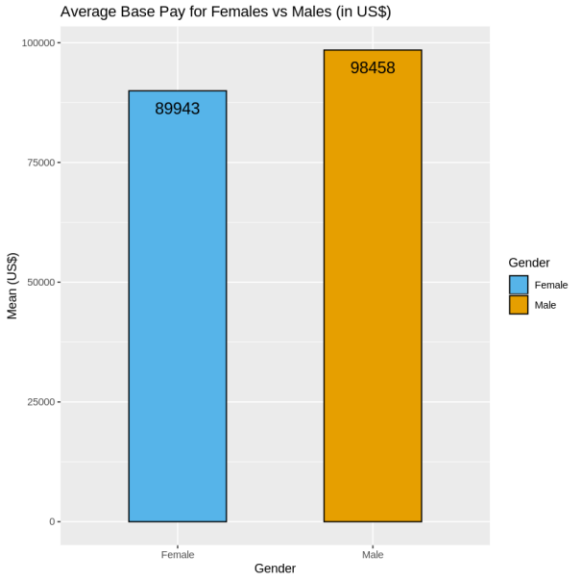



*Figure 7: Bar plot highlighting the mean difference in Base Pay*

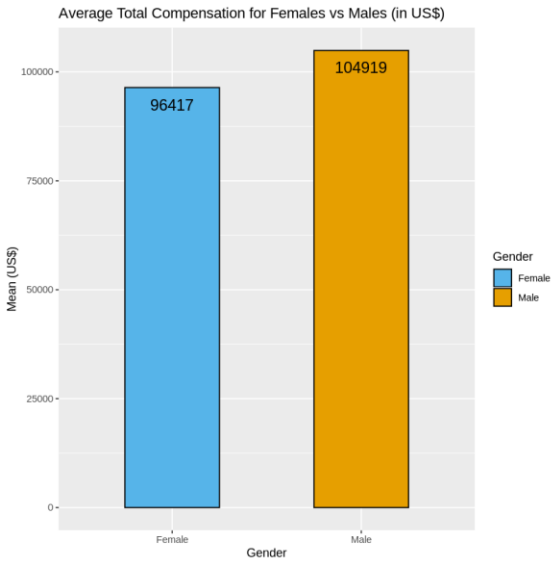



*Figure 8: Bar plot highlighting the mean difference in Total Pay*

## Calculating Pay Gap by Job Title

Figure 9 on the right shows the Pay gap between Males and females for every Job. Male Software Engineers make almost $11,886 more than female Software Engineers, which is surprising because software engineering is considered one of the more recent phenomena of the workforce, and many of our current technological advancements come from Females, female Data Scientists on the other hand make $7,200 more than male Data Scientists, and this could be explained because data science has seen a very recent boom in the economy, as the world becomes more data-driven. We also see a big difference for male drivers, making almost $5,220 more, as well as male Marketing Associates, who make $5,910 more than what female marketing associates make. Female Warehouse associates make $6,720 more than males do, which could be because of the physical requirements for the job, as well as the scarcity of females in that field.

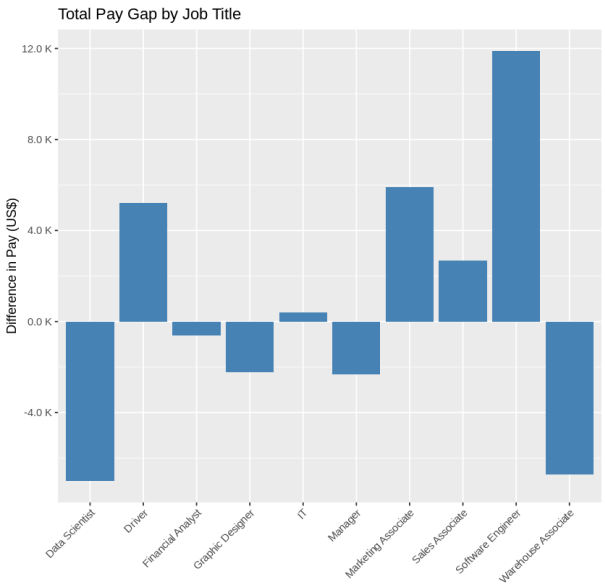



*Figure 9: Bar plot highlighting the difference in pay for every job*

## Regression Analysis

To finalize our statistical analysis, a **linear** regression model was built (shown in figure 10) to try to predict the trend between Male Pay and Female Pay. The values below are calculated from Figure 11.

$$Total\ Pay\ in\ Females = 5285 + 0.9409(Total\ Pay\ in\ Males) + \epsilon_i$$

The $r^2\ value\ is\ = 0.79\ (shown\ in\ Figure\ 11)$ meaning that 79% of the data fits the regression model. An r squared value of 79% indicates that the model explains 79% of the variability of the response data round its mean, meaning that in the future, as male pay goes up, female pay is also expected to go up, almost the same amount as Males (0.9409).

The Linear model also highlights the Job Titles, where we can easily see influential data points, high leverage data points, and also outliers in the data, for example, as we say above in Figure 9, Male Software Engineers make significantly more than Female Software Engineers, hence why the point is an outlier **and** point of influence, as the black curve highlights. on the data, since it is extreme on the x-axis. Since Managers earn the most, the point is shown to have very high leverage on the data. The p-value shown in Figure 11 is equal to 0.00054 (<0.05), which aligns with the hypotheses that Male Pay does affect Female Pay as well. Since p < 0.05 and r squared = 0.79, the regression model strongly suggests a positive correlation between Female Pay and Male Pay.

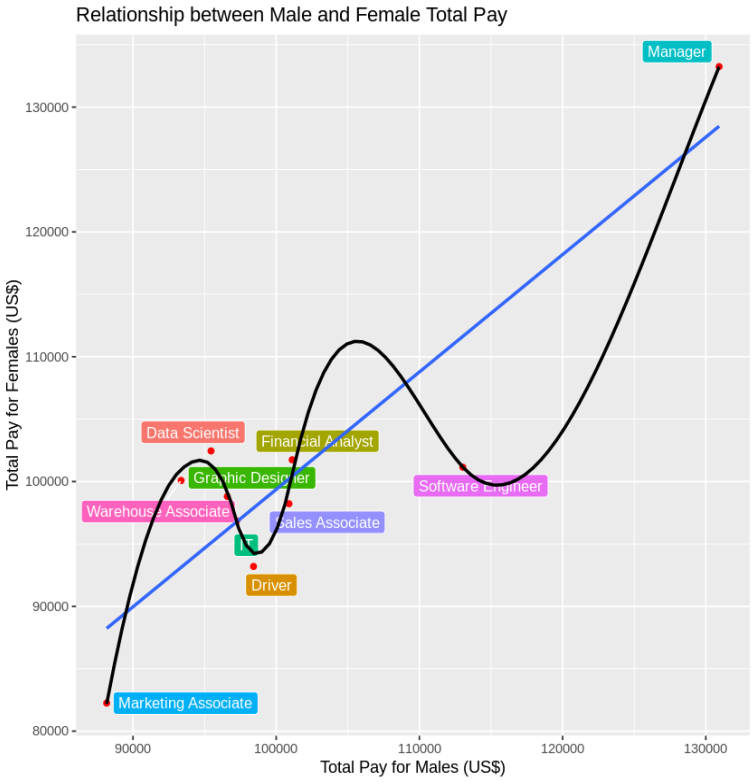



*Figure 10: Linear Regression Model also highlighting different points by Job Title*

```
Call:
lm(formula = TotalPay.y ~ TotalPay.x, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-10485.3  -4010.3    694.4   4244.7   7363.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.285e+03  1.731e+04   0.305 0.767911
TotalPay.x  9.409e-01  1.694e-01   5.554 0.000539 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
1 ' ' 1

Residual standard error: 6183 on 8 degrees of freedom
Multiple R-squared:  0.7941,    Adjusted R-squared:  0.
7683
F-statistic: 30.85 on 1 and 8 DF,  p-value: 0.0005385
```

*Figure 11: Summary of the Linear Regression model shown in Figure 10*

# Conclusion

Using various statistical analyses and data visualization, interesting discoveries and significant results were found. It was found that although in certain fields like Software Engineering and Marketing, male employees earn more than female employees, the wage gap has decreased significantly over the years, resulting in some jobs yielding a higher salary for female employees like Data Science and Warehouse Assistants. After analyzing pay based on Job title, a linear regression model was built to unravel the relationship between male and female pay. The regression model resulted in a strong positive correlation between male pay and female pay, even predicting that female salaries will almost match male salaries in the future in other jobs as well. These results are significant because 10 or 15 years ago it would not be possible, without the advancements of society, education and technology.

# Limitations

The data was acquired from 1,000 employees reporting their salaries on Glassdoor. This is the biggest data set that includes salaries, age, seniority and Job Title that was found online. Salary Reporting is not something people prefer to make public, hence why not many data sets can be found publicly online. With more data, a more detailed analysis could be done based on Job Title, and possibly a more detailed inference could be assumed since there are only about 100 or so entries per job. This data set is also limited to the United States, a broader data set consisting of data from multiple countries could help conclude if the wage gap is more severe in other countries.

# References


1. Lindstaedt, F. (2019, January 7). *Gender Pay Gap among Data Scientists on Kaggle - Towards Data Science*. Medium. https://towardsdatascience.com/gender-pay-gap-among-data-scientists-on-kaggle-87b393aa21fe

2. *Introducing dplyr*. (2014, January 17). RStudio Blog. https://blog.rstudio.com/2014/01/17/introducing-dplyr/

3. *Data visualization with ggplot2*. Data Visualization with Ggplot2. https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html