Analyzing Gender Biases in Black Saber Software

Evidence Found in the Salary, Promotional, and Hiring Processes

Report prepared for Black Saber Software by Brownwall Consulting

2021-04-21

Contents

Executive Summary

Background & Aim

A company should always keep fairness and equality at the top of its priority list as biases based on things like race or gender seriously damage its integrity and efficiency. The issue with this, however, is that determining where exactly biases develop within a company can be extremely difficult. Often, biases can be hidden behind many layers of a company's processes, and thus become easily ignored. In this report, we are attempting to identify some of these issues within the Black Saber Software company. More specifically, we want to determine; if there is a significant income gap between male and female employees of similar status, whether there is an association between the gender of an employee and the number of promotions they receive, and any possible points of gender based bias along the hiring pipeline.

Key Findings

To summarise the results we have found:

- Men make up the largest proportions of employees on the highest paying teams, and take home higher average salaries than women on almost every team. Moreover, within the same seniority men are seen taking home higher incomes again across almost all groups. as seen in **Figure 3.1.6**.
- After adjusting for productivity and working time, women are shown to earn on average, over $2816 less than their male counterparts.
- Women appear to be judged on a slightly harsher level than men in the company, being the only receivers of a "Needs improvement" leadership review.
- In the promotional process, men are seen earning consistently more promotions than women, increasing as the number of promotions rises.
- After adjusting for working time and productivity, women are observed earning about 28% fewer promotions than men, suggesting a greater ability for men to earn a promotion in the company.
- Men significantly outnumber women in the number of applicants hired even though the initial number of male and female applicants was relatively the same. As seen in **Figure 3.3.7**.
- It was found that men scored significantly higher than women in AI graded areas where gender was easily distinguishable, such as speaking skills and leadership presence. Men scored on average 18% and 8% higher than women in these areas respectively. This reduced the number of women that passed phases 2 and 3 significantly.

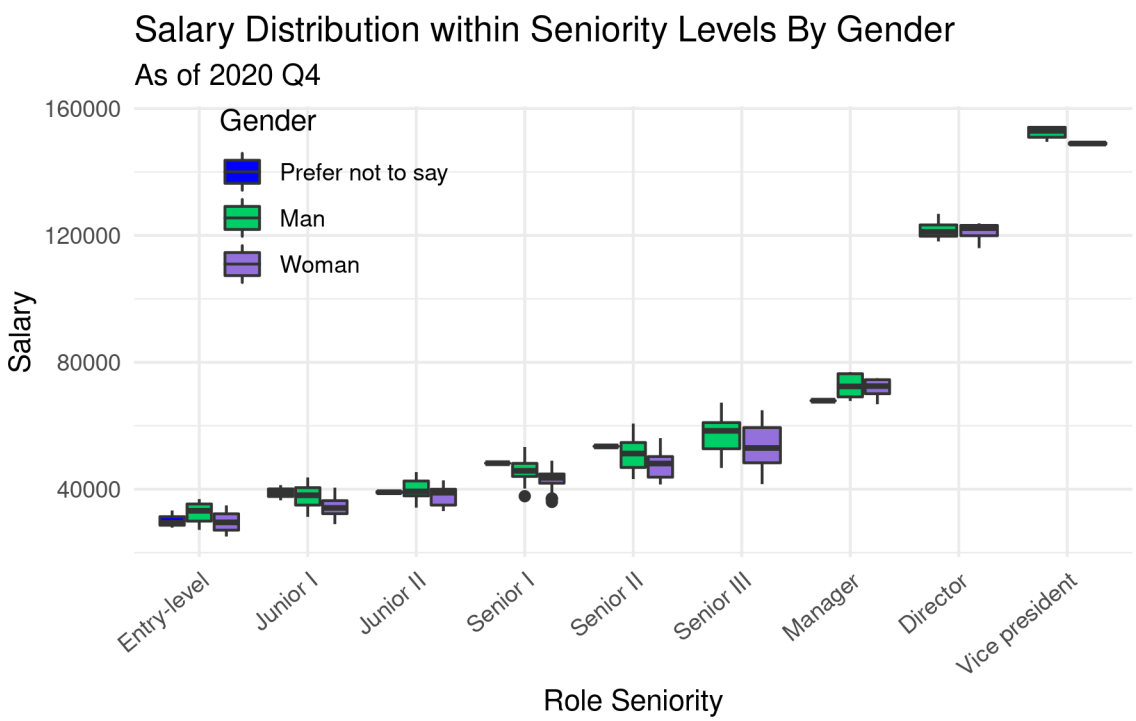

Figure 3.1.6: Black Center Line = Median Salary of Group

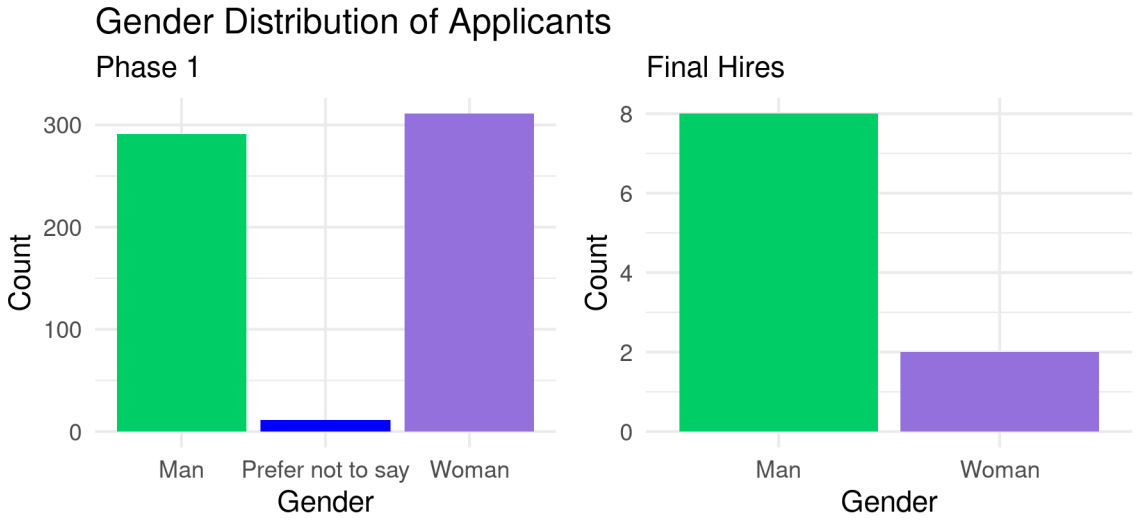

Figure 3.3.7

Limitations

- The strongest limitations of this study came from a lack of complete information.
- Gaining access to additional employee and AI data would allow for much more confident conclusions, and a more exact reason for the biases we observed.

Technical Report

1 Introduction

Biases in the workplace can be extremely problematic, especially when they are based on factors such as race or gender, as they put unnecessary ceilings above individuals that are working just as hard. Not only does this harm the company itself by creating needless friction between these groups, but it adds to the ongoing crises seen today by furthering discriminatory stereotypes and opinions. On top of this, biases can stem from many different areas within a company that are sometimes easy to turn a blind eye to. Thus, addressing and eliminating possible areas of bias should be a primary focal point for a company to both ensure the wellbeing of its employees, and the efficiency of work carried out.

In this paper, we are assessing the Black Saber Software company for possible areas of gender based biases within their hiring, promotional, and salary processes. Specifically we want to find answers to the questions; *is there a gender pay gap*, *do men tend to earn more promotions than women*, and lastly, *is the hiring pipeline fair in its assessment of male and female applicants*. This was accomplished thanks to the large and detailed datasets provided by Black Saber's data team. Answering these questions involved conducting many statistical analyses on these datasets and looking for associations between any of these factors and an individual's gender.

Briefly, we found evidence of gender bias of varying degrees in all three of these sectors. In terms of salaries, this largely stems from men being favoured for the more technical and high paying teams. The promotional bias seems to come from an internal resistance to placing women in higher ranking positions. And lastly, the AI in the hiring process appears heavily biased in its assessment of applicants, that is then passed on to the final hiring decisions.

To outline the paper, firstly, in *Section 2* we provide some explanations on the data used for the analysis, as well as some initial manipulations made accordingly. In *Section 3* we examine each of our research questions in depth and relay and interpret the results we find. *Section 4* then provides a discussion of these results as well as the strength and weaknesses of the analysis, and areas for future work. Finally, *Section 5* contains some consultant information for our company including consultant profiles for the analysts working on this report, as well as an outline of the code of ethical conduct we follow within the company.

2 Data Descriptions

2.1 Employment Data

To understand both the salary and promotions processes, and any biases behind them, the dataset provided by the data team on the current employees at the company was used. This dataset featured observations for every employee at the company during each financial quarter they were actively working in, measuring 8 different variables. Information on the employee's; id, gender, team, seniority, and salary, as well as their productivity and leadership ratings, was included at each observation. Employee Id was a simply 5 digit variable, naturally matching an employee's ID number. Gender was a qualitative variable with three available options; "Man", "Woman", and "Prefer not to say". Team was a categorical variable with 8 different options; Client Services, Data, Design, Legal and Financial, Marketing and Sales, Operations, People and Talent, and Software. Role seniority measures the relative position and status of an employee in the company, again with 8 categories with increasing importance as follows; Entry-level, Junior I, Junior II, Senior I, Senior II, Senior III, Manager, Director, Vice president. Salary was simply a variable that wrote out the salary of an employee in a standard form (i.e. $10,000). The financial quarter variable represents the quarter in which the observation was taken, starting from the second quarter of 2013, ending in the fourth quarter of 2020. Lastly, the ratings variables were determined by the higher executives in the company. Productivity is graded on a scale from 0-100, with 50 being the satisfactory level, where 50+ is considered better than expected. And leadership was rated based on the employees' role as; "Appropriate for level", "Exceeds expectations", and "Needs Improvement". This dataset contained a total of 6906 observations across all 607 unique employees.

Employment Data: Wrangling

Before diving in deeper to analyse the research questions, a few initial modifications were made to the data that would allow for better statistical interpretations. Firstly, as the initial salary data was in character form, these numbers were parsed out to obtain their numerical values. Then, an additional column was created that tracked the number of financial quarters an individual had been employed for. This was done simply by taking a cumulative sum of the number of individual observations for each employee, ordered by their financial quarter. As the length of a person's employment is likely to have a strong effect on both the promotions they receive and the income they earn, this variable was extremely useful to the analysis. By measuring the number of changes in salaries of the employees, a column was created that calculated this total at each respective observation. After some more manual inspection, it was found that all of these salary changes were increases, with no signs of any decreases. This then allowed for us to track which

individuals received the most raises throughout their employment history, as well as its trends with other variables. In a similar manner to the manipulation made for salary changes, another column was added that tracked the number of promotions an individual employee has earned. This was determined by identifying the point at which an employee's role seniority changes, and summing together the number of these occurrences. As a result, we were able to study which employees received the most promotions and whether or not this result was associated with the employee's gender.

2.2 Hiring Data

Next, using the multiple hiring data sets, we will be analyzing Black Saber's hiring pipeline and new AI software system to determine if the process used to select new hires is biased towards men. The process for hiring is broken down into 3 phases and includes 4 data sets; phase 1, phase 2, phase 3, and final hires. Phase 1 was the initial application phase where applicants were removed if they did not pass certain baseline requirements set by the AI system. This contained information on all 613 of the initial applicants. The data set for Phase 1 specifies; which team applicants were applying to, Data or Software, if they provided a cover letter and/or a CV, both graded binarily as 0 or 1, their cumulative GPA on a scale of 0.0 to 4.0, gender, extracurriculars, and work experience. Gender had 3 available options; "Man", "Woman" and "Prefer not to say", from which applicants could select. Extracurriculars were graded by the AI on a scale from 0 to 2, where the level is determined by the amount of involvement, and the relevance and usefulness of the skills gained from these opportunities. Similarly, work experience was also graded on a scale of 0 to 2 and was determined using relevant information from the application such as company names and reputations. For this scenario, 300 out of the 613 individuals who applied for a job at Black Saber made it past phase 1. Next, for applicants to make it past Phase 2, they were required to finish both a timed technical and writing task and submit pre-recorded video. In this phase, Ai was used to measure applicants writing skills, technical skills, speaking skills and leadership presence. Technical skills and writing skills were graded on a scale from 0 to 100 by AI using the applicants' timed technical and writing tasks respectively. Speaking skills and leadership presence were scored by AI, on a scale from 0 to 10, using the pre-recorded video applicants submitted. In order to compare the effect of these factors on whether an applicant passed phase 2, we multiplied both speaking skills and leadership presence by 10 so all of the factors are ranked on a scale from 0 to 100. AI then uses these scores to filter out and narrow down the applicant range. In this study, total applicants were narrowed down to 22 in phase 2 from 300. Finally, Phase 3 was the interview stage. Applicants that made it past the first 2 phases had to go through a series of 2 interviews in which they are given an overall rating from 0 to 100 for each interview on how well they fit the job. Results from the interview are then used to determine which applicants got hired. At total of 10 out of the 22 applicants that made it to

phase 3 were hired in this example, 5 for the software team and 5 to the data team.

Hiring Data: Wrangling

In order to both fully visualize and model the data for this study, 4 additional datasets were created from the original 4 hiring data sets. First, a dataset that combines the data from all 4 datasets was created in order to fully visualize the data and to use in the first regression model. Additional columns displayed in Phase 2 and Phase 3 were added to the Phase one data set to create this dataset. In addition 3 columns were created in order to signify if an applicant had made it to that phase or has been hired. Next a dataset which combined data from phase 2 and 3 was created in order to model which factors in phase 2 affected if an applicant were to pass the phase. Another dataset which combined data from phase 3 and final hires was created to use in the model which determines what factors affect interview ratings. A new column was created, which displayed the mean interview rating applicants received. Finally, one last dataset was created which contained information from all phases but only included applicants that were hired in order to visualize gender distributions of final hires.

3 Research Questions

As explained, this study covers three primary research questions;

- Is there a Gender Pay Gap?
- Do men have an easier time earning promotions than women?
- Is the hiring pipeline fair in its assessment of male and female applicants?
- Note: While our research questions primarily target male vs. female differences, we have still included individuals that did not identify their gender, both out of respect to these individuals, and for a full scope of the data.

3.1 Analyzing Salary Processes

Methods

The statistical model used to answer this question was a linear mixed model on the new numerical salary value, for all 607 employee observations from the most recent quarter. This featured covariates for an employee's; gender, productivity score, and working time, with a random effect for an individual's role seniority. This model was chosen primarily due to the relationship observed between salary and role seniority, as displayed below in **Figure 3.1.1**.

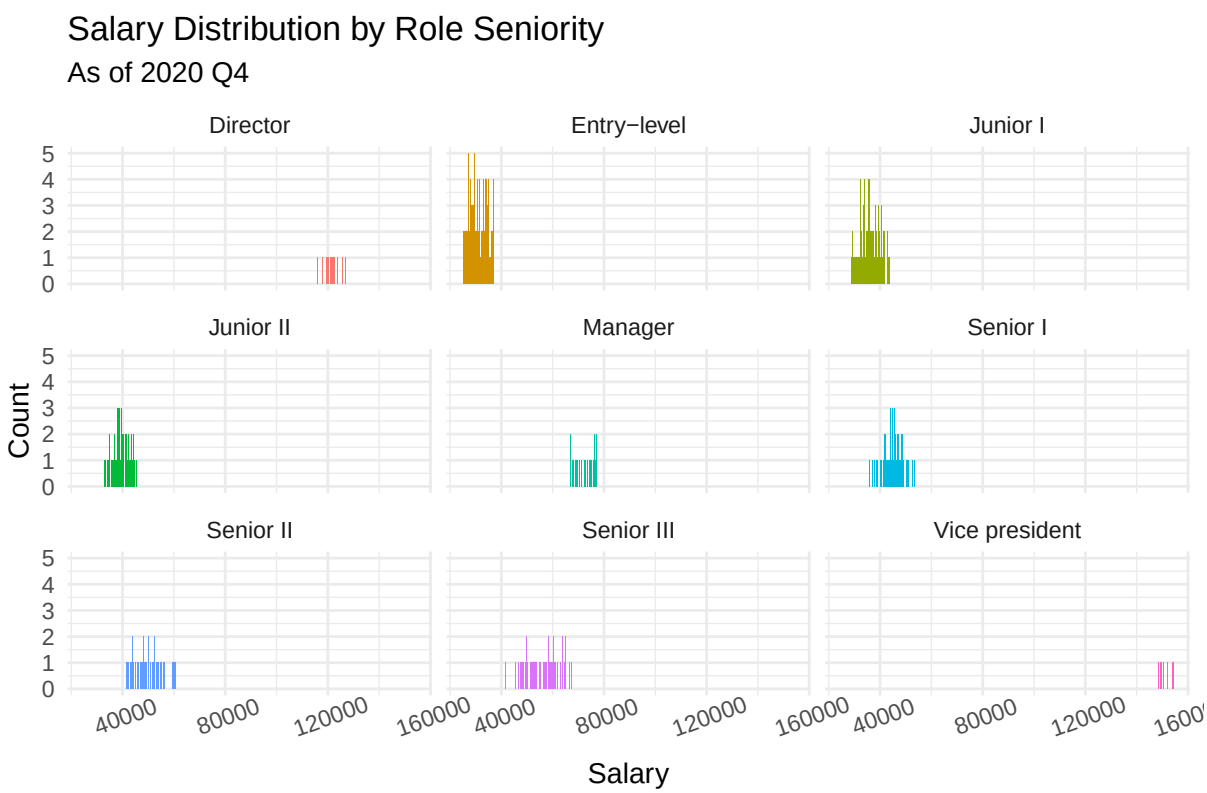

Figure 3.1.1

As the salaries within each role appear quite close together, and somewhat evenly distributed in these groups, a linear mixed model stood out as a solid choice. The variables used in this model were chosen due to the importance we felt they would bring, as well as the ability to isolate more of gender's specific effects on salary. Since we are only analyzing employee observations from the most recent quarter, there is only one per employee, which makes us much more comfortable confirming that the independence assumption has been satisfied. Thus a linear mixed model was used for this analysis.

Results

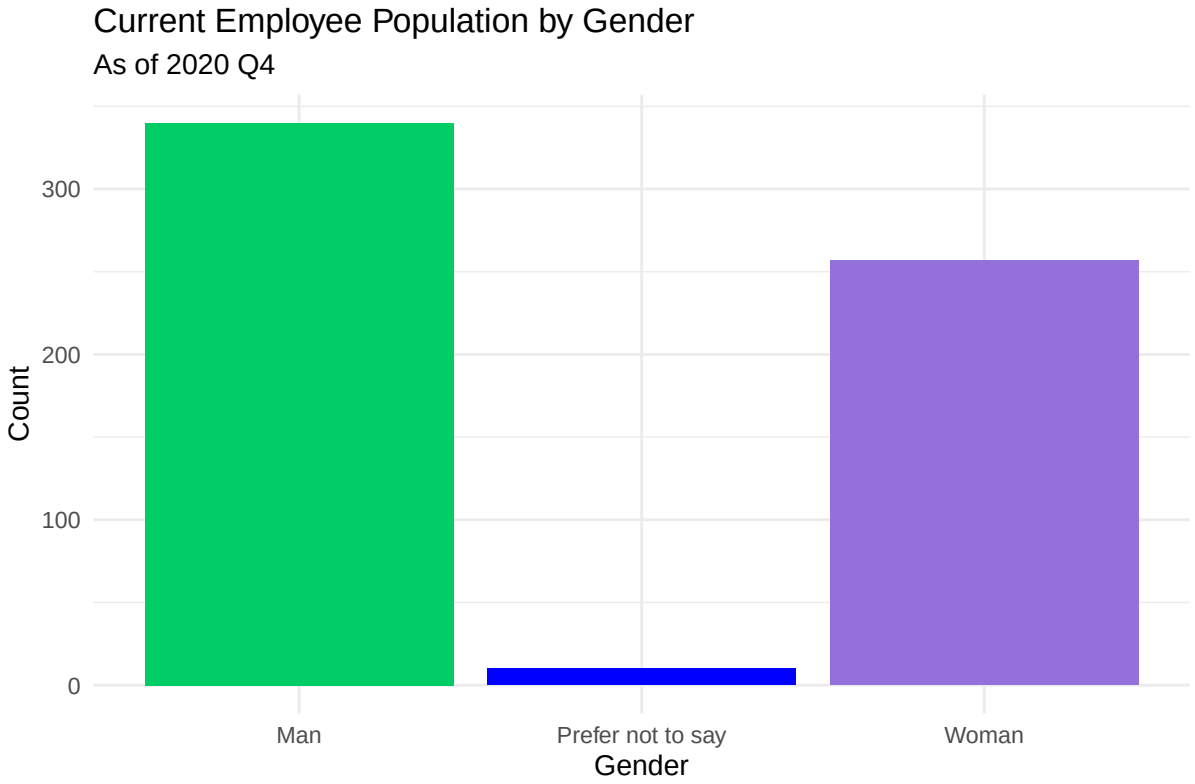

Figure 3.1.2

To begin with, we have found that there is a significant difference between the number of women and men employed at the company, as displayed in **Figure 3.1.2**. To be exact there are 83 more men than women currently working as of the most recent financial quarter. While this is not definitive proof of a bias necessarily, it is still a large enough difference to draw some concern given the relatively small pool of employees, and thus require further inspection.

Upon further analysis, we can take a look at the distribution of genders both within teams and within roles as of the most recent quarter.

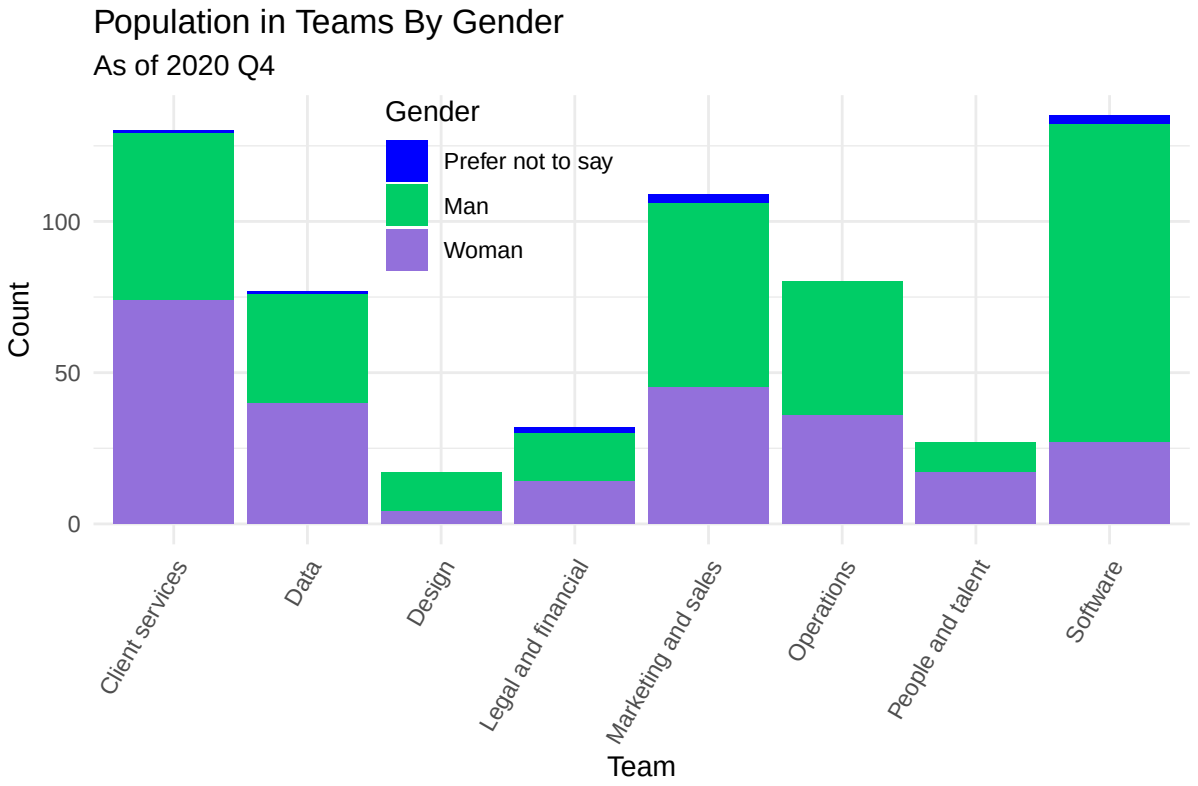

Figure 3.1.3

Here, in **Figure 3.1.3**, we can see that men are dominant in many teams including; software, marketing and sales, as well as the design team, whereas women appear to make up a majority of the client services and the people and talent teams. The remaining teams, however, have relatively similar proportions of men and women with only slight differences that we might regularly expect to see This result implies that men are favoured to carry out tasks seen as more analytical or computational, whereas women are favoured for the personal relations and communications side of the company. It is difficult to determine, however, where along the hiring process, this bias exists, as we would need to be supplied with the hiring data of the current employees list. While it may seem easy enough to make this reasoning, basing these hiring decisions off of a person's gender is a serious and oppressive problem. As it unnecessarily restricts their future both within the company, as well as in other areas of work, gender should never be a considered factor in the field an employee fits into.

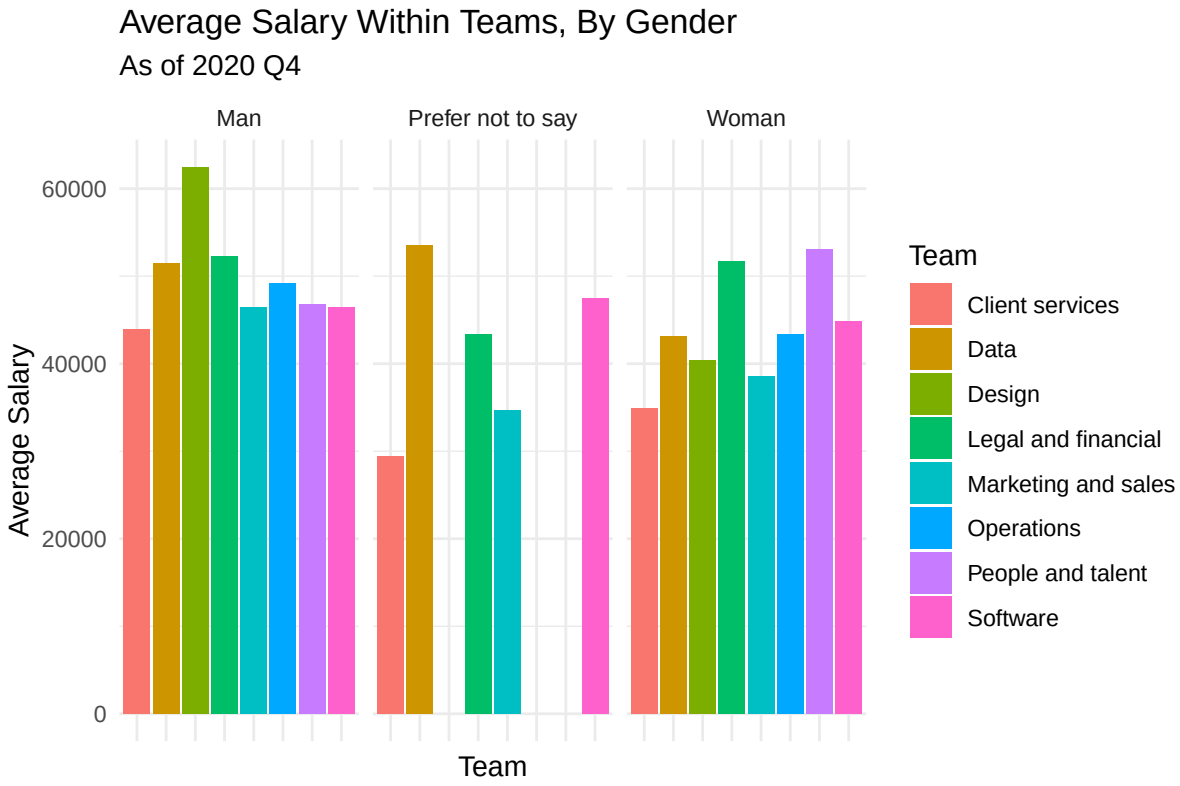

Figure 3.1.4

To understand how this result affects the salary of employees we can observe the distributions of average income earned by each gender within these teams, shown in **Figure 3.1.4**. As we can see, for almost every team, male employees earn a slightly higher average level of income than the female employees. Most notably, the design team, which takes home the largest incomes, pays a significantly higher amount of income to the male employees. This, coupled with the fact that it is already a male dominated team, signifies a possible favouring towards men when it comes to salaries. The only team on which women show better incomes is the people and talent team, but by making up such a small percentage of the working population, it pales in comparison to the categories dominated by the male employees. This presents an initial indication of there being an association between the salaries of employees and their genders. More specifically, it shows that women within the same team as their male coworkers may earn less income, simply for being a woman. Still, however, additional analyses must be conducted to understand the other factors at play here and determine the true existence of a wage gap.

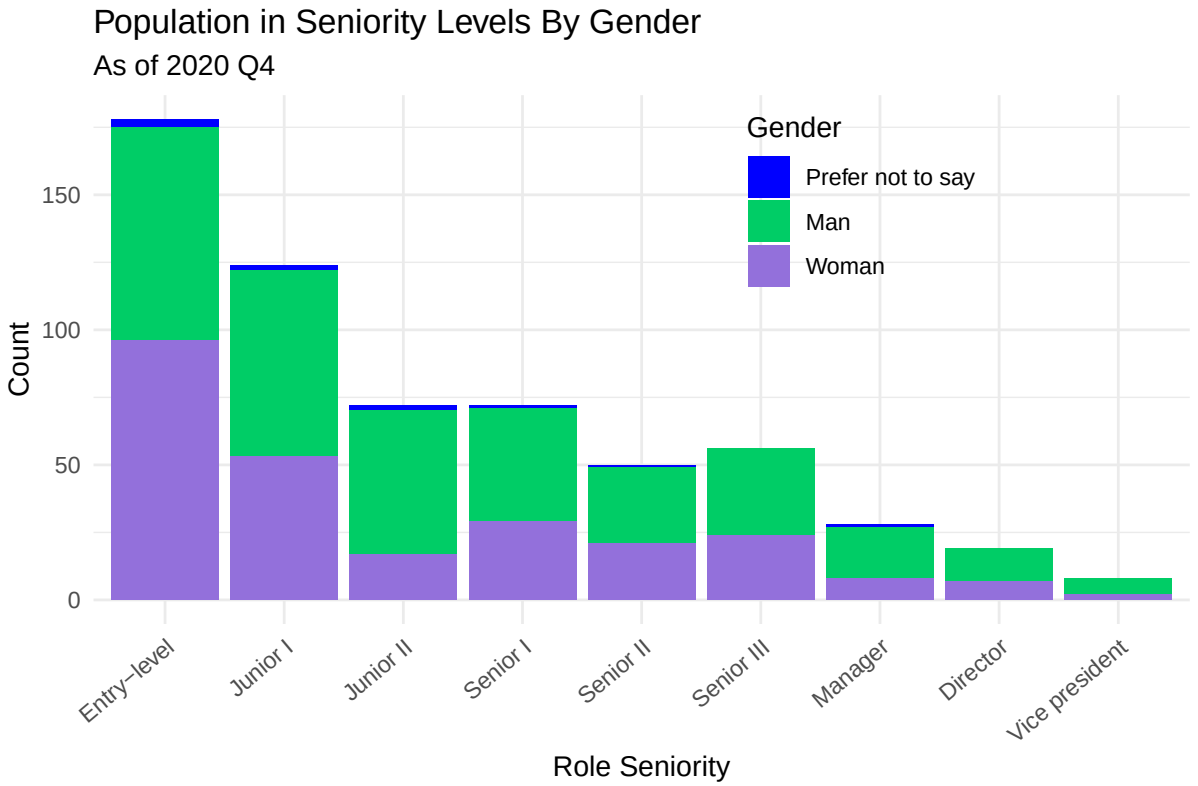

Figure 3.1.5

To further visualise this bias, the roles of employees were seen to have a significant effect on their incomes, thus the proportion of genders within each role has been displayed in **Figure 3.1.5**. As shown, for many of the lower level roles, like the junior and senior levels, there is a relatively even balance between men and women. However, in the more executive roles of director, manager, and vice president, men begin to show a slight advantage. This signifies possible evidence of an inherent resistance to placing women in higher positions of power within the company. While again more factors must be assessed to conclude on this bias, this result can be seen as highly oppressive and must seriously be reviewed.

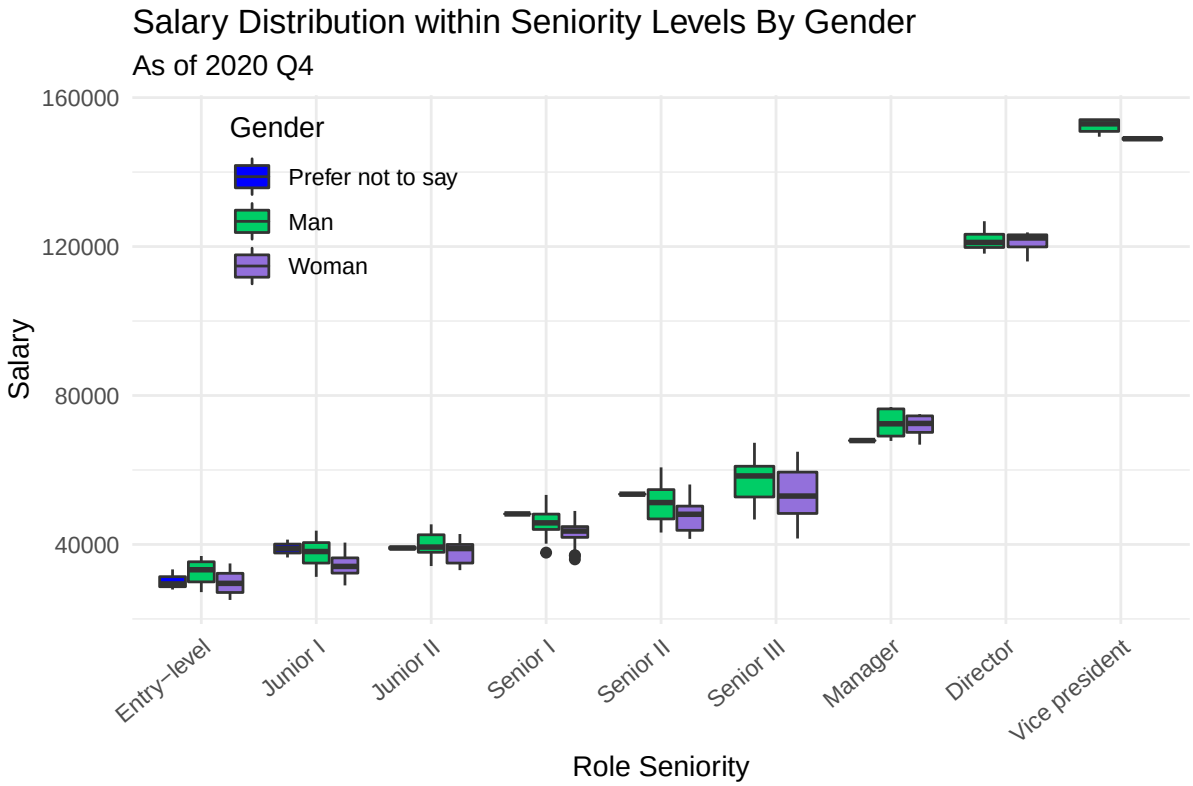

Figure 3.1.6: Black Center Line = Median Salary of Group

Moving to the salaries of these roles, seen in **Figure 3.1.6**, as expected, the more executive roles take home much higher incomes than the senior and below roles, regardless of gender. However, when we remember that these executive roles are largely male dominated, this result turns out to significantly favour the male employees of the company. When we consider the split between genders, and study the salary distributions, we can see that across almost all roles, men tend to take home a greater income than their female counterparts, with higher median incomes at all seniorities except the Director level. In a similar vein to the issue seen in the teams of employees, this is a result to be alarmed by as it signifies a level of inequality within the salaries of employees. Thus, again, we can see that there may be a slight bias toward men in the company, suggesting the possible existence of a gender pay gap.

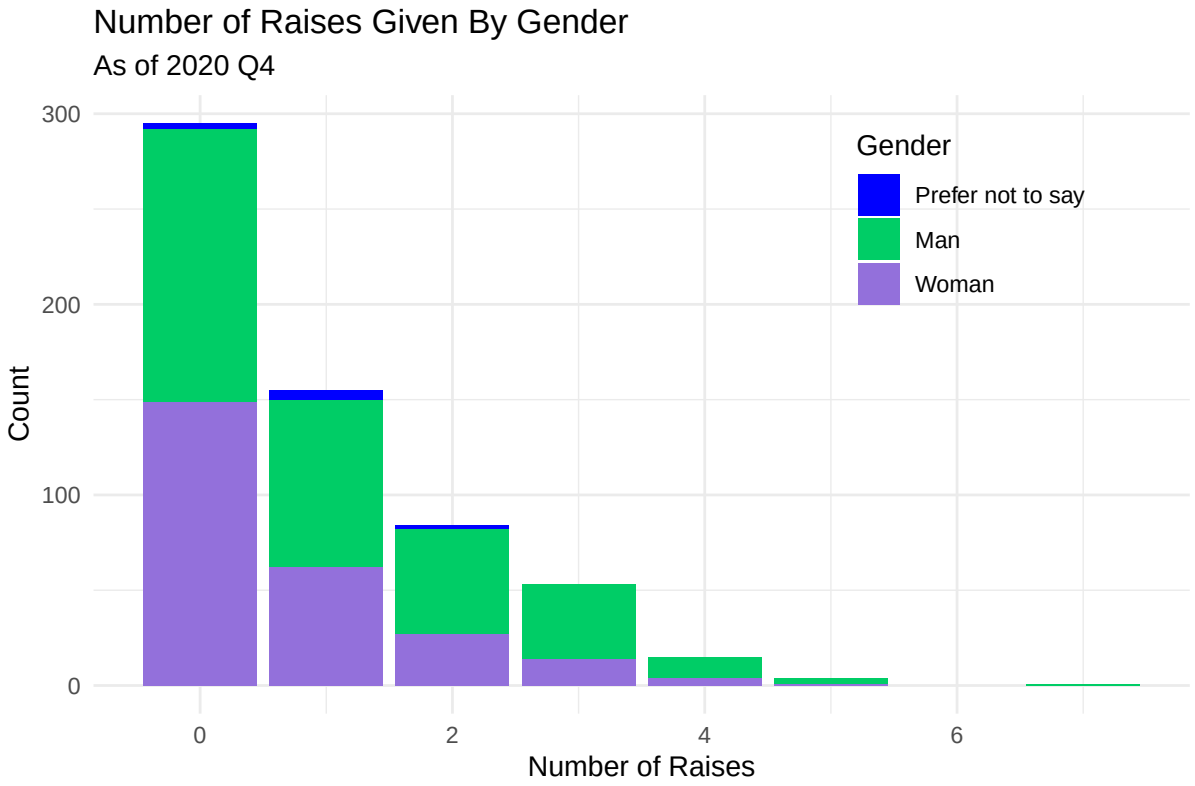

Figure 3.1.7

In addition to the salaries currently earned by the employees, we can also study the number of salary increases each employee receives, to observe whether or not it has an association with gender. To do this we have created a bar chart, in **Figure 3.1.7**, that plots the number of employees that have received each number of raises, coloured by gender. As displayed, a relatively even distribution is seen for those with no salary raises, but this quickly becomes increasingly more male dominated as the number of raises rises. Thus, we find that the company may be favouring the male employees when it comes to deciding on who should receive a raise, which can in turn fuel the wage gap.

Table 1: Linear Model Output

	Estimate	Std. Error	t value
(Intercept)	68484.53	13927.13	4.92
genderPrefer not to say	-776.36	1174.67	-0.66
genderWoman	-2816.11	308.72	-9.12
working_time	9.30	42.02	0.22

	Estimate	Std. Error	t value
productivity	-11.55	10.29	-1.12

Table 2: Confidence Intervals

	2.5 %	97.5 %
.sig01	26328.69	67859.60
.sigma	3439.89	3852.91
(Intercept)	39762.55	97222.79
genderPrefer not to say	-3075.15	1521.44
genderWoman	-3420.30	-2212.26
working_time	-72.75	91.71
productivity	-31.69	8.58

Finally, to gain a thorough and numerical understanding of these possible biases, a linear mixed model was applied to the full employment data to observe some covariate effects on the incomes of employees. This included the gender of the employee, the number of quarters the employee has worked, and their productivity score, as well as a random effect term for the id and seniority of the employee. The output from the model, along with the 95% confidence intervals for the estimates are displayed in **Table 1** and **Table 2**, respectively. Most importantly, this model estimated that, after controlling for productivity and working time, women earn *$2816.11* less than a similar male employee in the company. The 95% confidence interval for this estimate predicts this value to be between *$3420.30* and *$2212.26*, thus as it does not pass over zero, we can safely reject the null hypothesis that this factor's true effect on employee salary is zero. In turn, this is highly significant of a possible gender bias that exists within the company's salary process, that works to benefit the men more than the women. The remaining factors, productivity, and working time, surprisingly showed little significance to the model as their confidence intervals both included 0, preventing the rejection of their null hypotheses and any subsequent conclusions. This adds more strength behind the association between an employee's salary and their gender as it stands out as the only strong predictor.

Thus, we have found some significant evidence of a bias existing within the salary processes at the company that creates a wage gap, which must be reviewed and ideally resolved.

3.2 The Promotional Process

Methods

To study the promotional process in the company, along with multiple visualizations, a zero inflated Poisson model was fit to the current employee data from the most recent quarter, using the column measuring the number of received promotions as its response. The variables included in this model were; gender, productivity, and working time, to then determine the excess zeros, a new variable was created that indicates whether or not an employee is at entry level seniority. This choice was made primarily due to both the response type and distribution within seniorities. Firstly, as the number of promotions is a count variable, Poisson regression is an ideal option as it works best with this data type. However, upon inspecting the distribution of this response, in **Figure 3.2.1**, we find an extremely high number of zeros, pushing us toward a zero inflated model.

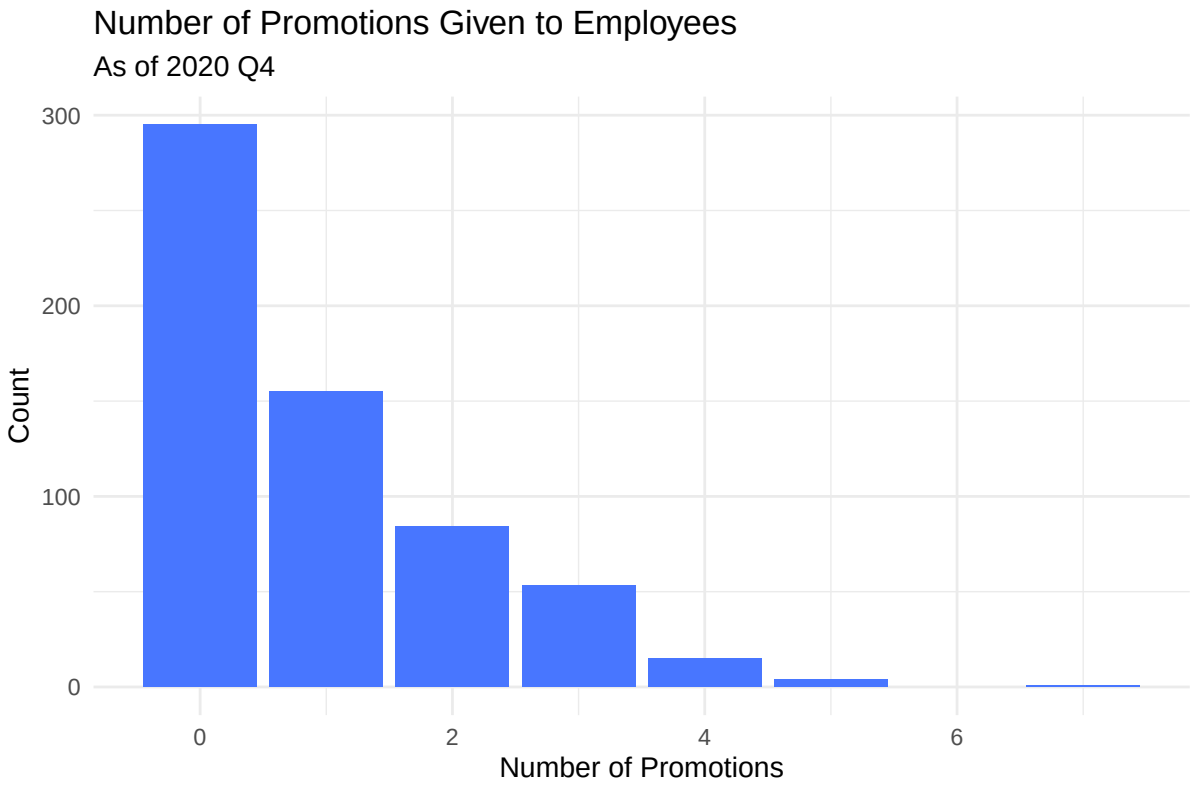

Figure 3.2.1

Taking this one step further we examined the specific distribution of promotions received in each seniority group, as displayed in **Figure 3.2.2**.

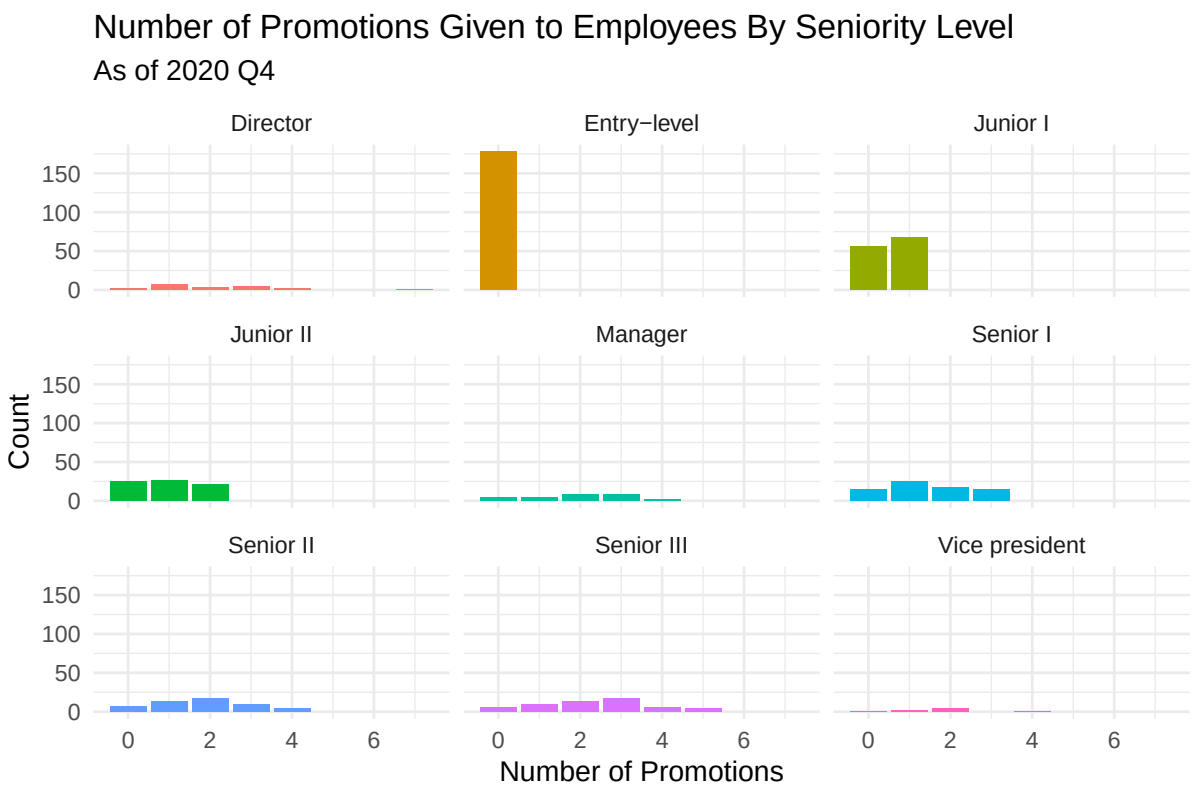

Figure 3.2.2

As we can see, quite naturally, everyone at an entry level position has yet to receive any promotions, suggesting a separate distribution from which the excess zeros come from. Unfortunately, there is no easy method of assessing overdispersion within a zero-inflated Poisson model. However, when conducting a Vuong test comparing this zero inflated model, to a similarly designed Poisson model, we saw that it performed significantly better. Lastly, since we have fit this to data only from the most recent quarter, we are fairly comfortable assuming that each observation is independent of the other as we only generate one observation per employee. Moreover, the only other area of dependence we considered was within teams, however, examining the promotions within these teams showed a very similar distribution across them, suggesting no significant relationship between those on the same team. Thus, we believed that a zero-inflated Poisson model was the best option available to us.

There is some concern, however, with the exact way the model will handle this overdispersed zero group, but with the given information we felt that again, it was our best option.

Results

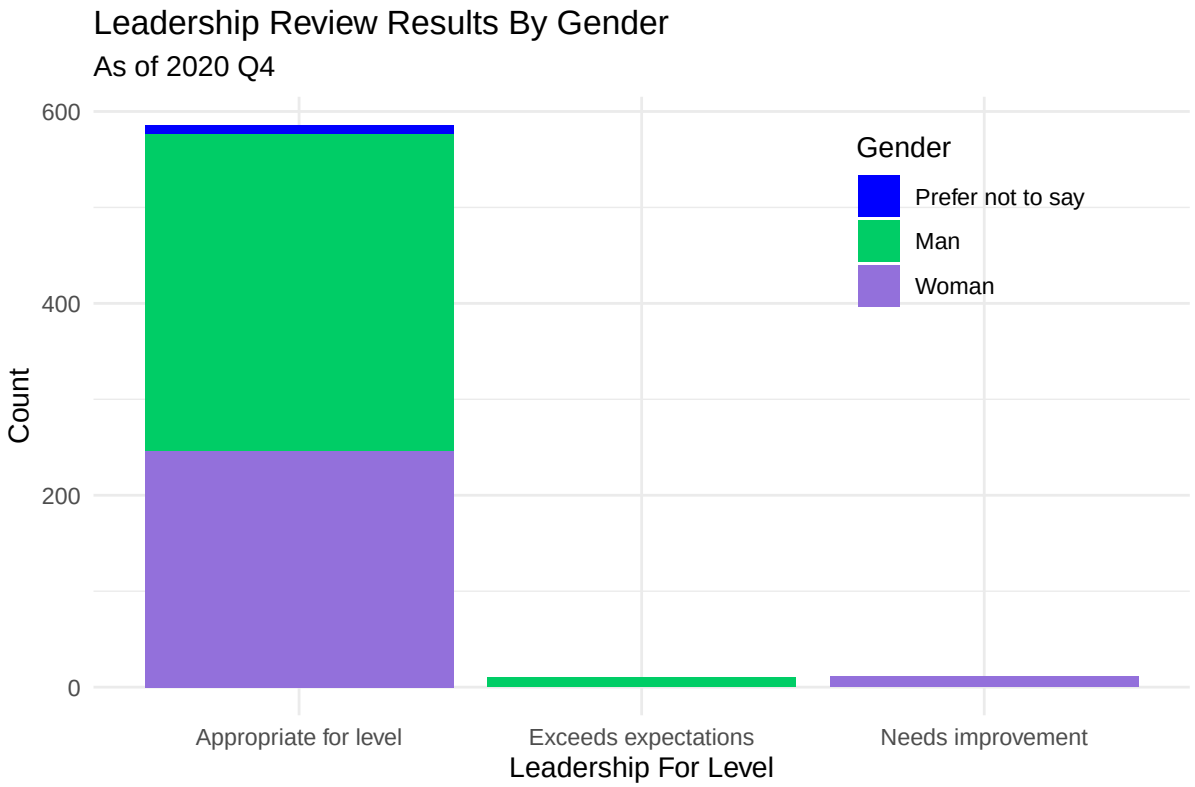

Figure 3.2.3

In the evaluations of employees at the company, the leadership reviews for each role show a serious issue with the differences between how men and women are regarded. In **Figure 3.2.3** we see that, specifically in terms of these reviews, women are the only ones given a "needs improvement" rating, while men are the only ones seen receiving an "exceeds expectations" rating. This suggests an initial higher standard that women may be held to in the company, that may be somewhat unrealistic for these individuals. Conversely, this shows that men may be more highly regarded in the company and thus given more favourable reviews of their performance, simply for being a male. That being said, the majority of these employees are shown to have a leadership ability that is appropriate for their specific level, which is by itself a fairly good sign of a relatively efficient working environment.

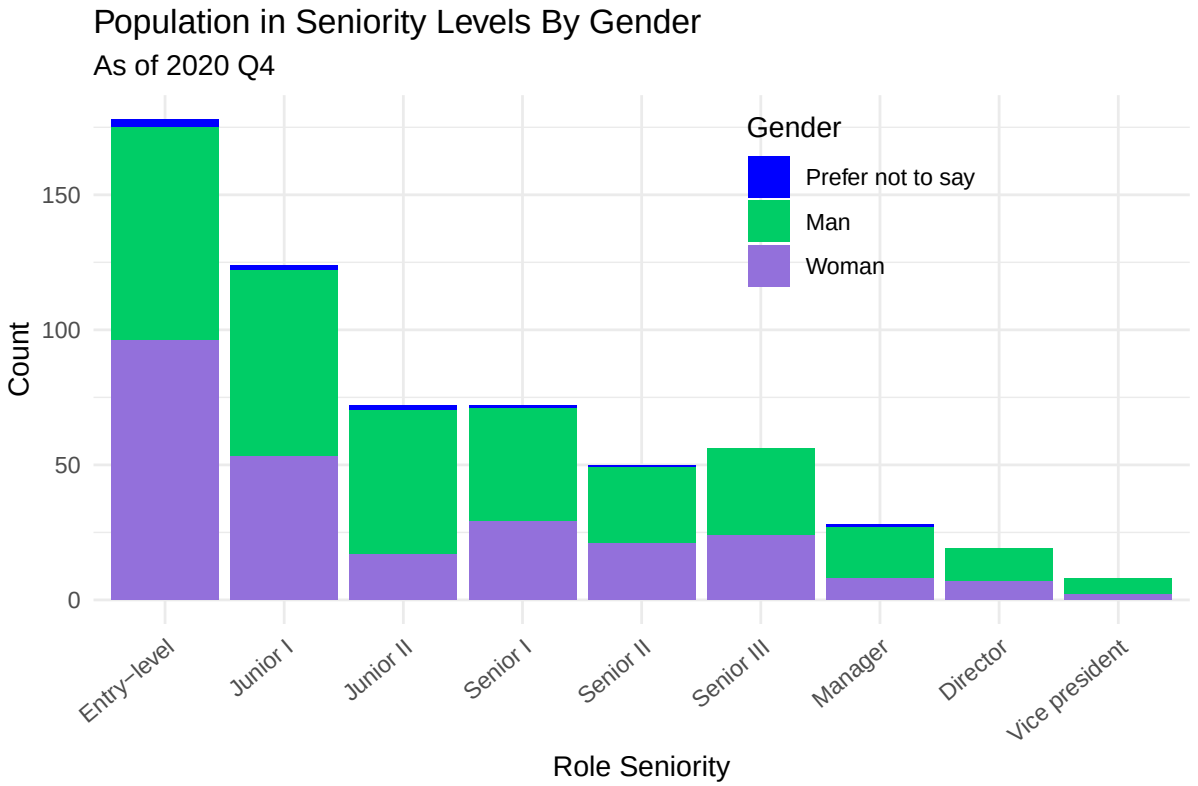

Figure 3.2.4

It is extremely important to also keep in mind that these reviews come from those in higher executive positions, with the greatest amount of power. When we remember from **Figure 3.2.4**, that most of these executive positions are filled by male employees, it becomes clear where this bias may be coming from. This is likely suggestive of the fact that men may be more inclined to side with other men rather than objectively analyze what is in front of them. However, this is only a speculation as we do not have information on exactly which sectors of the company are responsible for these reviews. While this figure does seem to suggest that men tend to receive more promotions overall, we cannot use it as such because it provides no information on any of these employees' start dates. This then prevents us from determining whether they were placed at that level to begin with, or received a number of promotions to reach the point they are at currently.

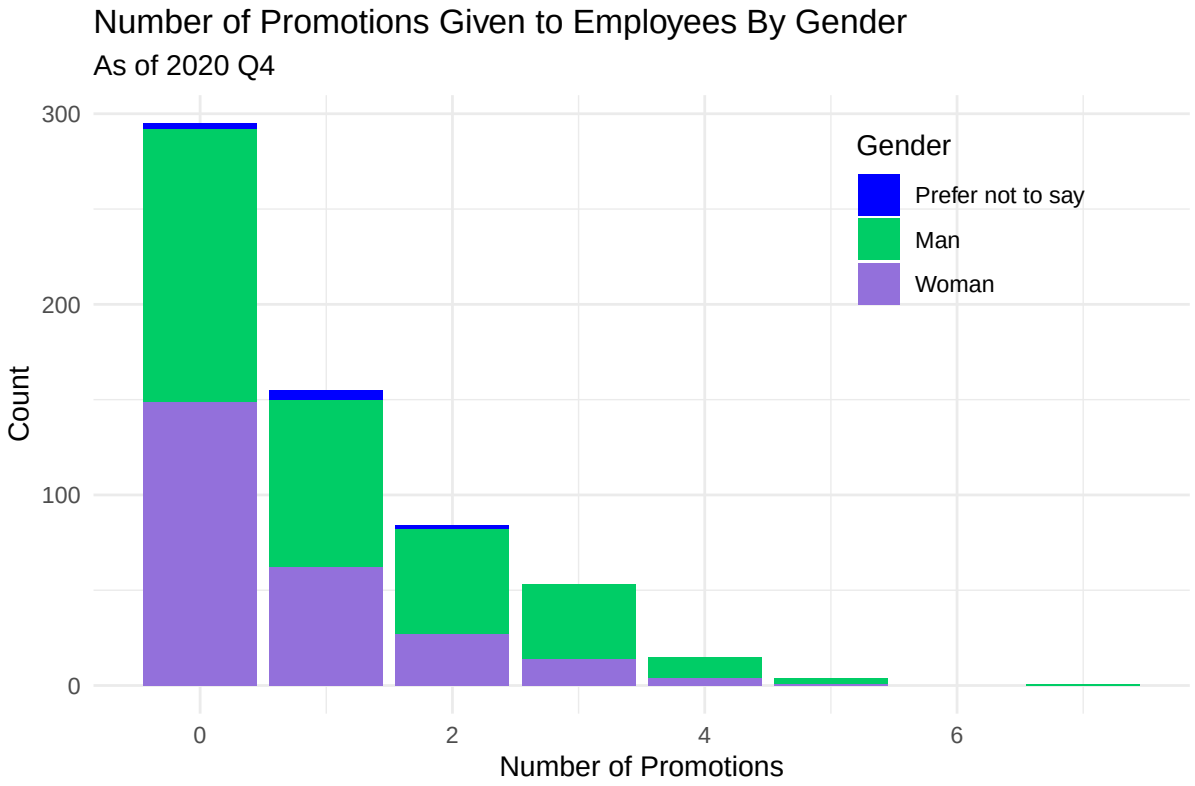

Figure 3.2.5

Now, when looking at the number of promotions each employee has received in **Figure 3.2.5**, we see a similar distribution to the plot of raises per employee. The greatest proportion of employees sit at the no promotion level, where there is an almost equal number of male and female employees. This is expected given the large amount of relatively new workers that have not yet qualified for a promotion. However, as we move up to higher numbers, men quickly take over and are seen as the main receivers of promotions, especially at the higher levels. This result is again, strongly suggestive of another area in the company where men may be favoured more than women, that must be carefully reviewed.

Table 3: Zero Inflated Model Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.349	1.218	0.005	1.000
genderPrefer not to say	0.868	1.403	0.658	1.965
genderWoman	0.726	1.095	0.030	1.000
working_time	1.094	1.007	831412.432	1.000

	Estimate	Std. Error	z value	Pr(>|z|)
productivity	1.001	1.003	1.438	2.047

Using a zero inflated Poisson model on the number of promotions, we have found some interesting results that explain the rates at which employees earn promotions. This model included covariates for gender, working time, and productivity, and it used an indicator for whether or not the employee was an entry level worker to map the extra distribution of zeros. After controlling for working time and productivity, from **Table 3** we can see that women earn significantly less than men in the company. Specifically, this model reports that women earn about *27.4%* fewer promotions than men that have worked a similar amount of time at a similar productivity level. With a 95% confidence interval that reports decreases between 13% and 40%, we are able to reject the null hypothesis that this factor has no effect on the number of promotions received. As a result, this signifies a strong bias against female employees that tends to give male employees more promotions, simply for being male. The other factor that showed a significant result was the working time of an employee. Naturally, this estimate suggested that those with longer working times are likely to have received a greater number of promotions. Specifically, for advancing 1 financial quarter, an employee is expected to receive about *9.4%* more promotions, controlling for the other factors.

To summarise, we have found a few instances of a slight promotional bias within the company. These create an environment wherein men are more likely to earn a greater number of promotions throughout their employment than women, for a relatively similar amount of work.

3.3 Hiring Pipeline Analysis

Methods

To interpret biases in the hiring pipeline, three different models were used at each succession point. Going from phase 1 to phase 2, we used a logistic regression based on the variable that indicates passing phase 1. Naturally, as this was a binary variable, logistic regression stood out as the most effective choice. Covariates for an applicant's gender, gpa, extracurricular activities, and work experience were added to this model. This was based on the relationship we found with these variables and the success of an applicant in phase 1.

Secondly, between phase 2 and phase 3, another logistic regression was fit on the response variable indicating a pass or fail in phase 2. Again, as this was a binary variable, logistic regression was determined to be the best possible option. This time, we added covariates for gender, as well

as all 4 test scores autograded by the AI; technical skills, writing skills, speaking skills, and leadership presence. This choice followed a similar reasoning to the first model as all of these covariates appeared to be significantly important in understanding any biases from the AI.

For both of these models, preliminary plots shows no significant sign of a grouping based on passing either phase within both teams. Thus we believed that our models also satisfied the independence assumption.

Finally, as passing phase 3 was found to be highly dependent on an applicant's interview scores, a linear model was fit using the mean of these interview ratings as its response.

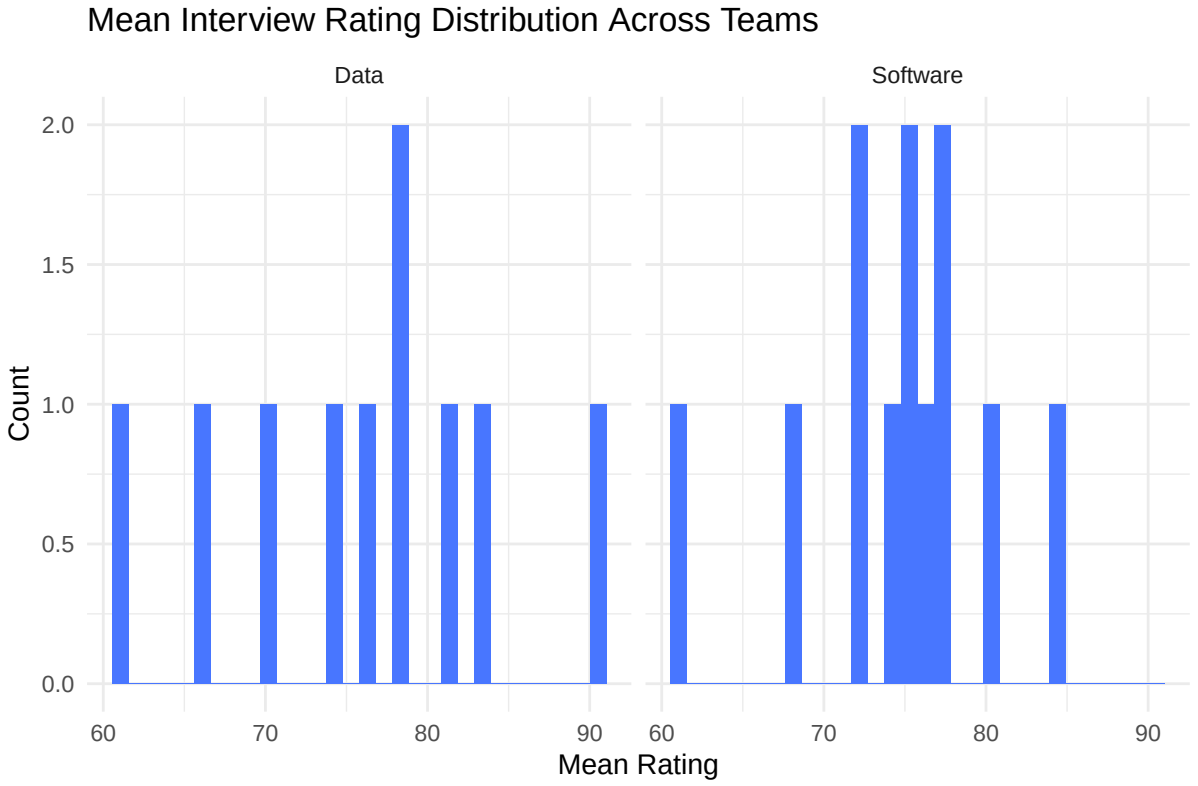

Figure 3.3.1

Studying the distribution of this mean rating, given in **Figure 3.3.1**, we found that, while it is not as continuous as we may like, it still shares a relatively even distribution around its center, and shows only very slight tail deviance in a qqplot. Additionally, we are relatively happy with assuming independence from these plots. Thus we remained fairly comfortable with our decision to use a linear model for this analysis.

Results

The hiring process was broken up into 3 main phases.

Phase 1 to Phase 2:

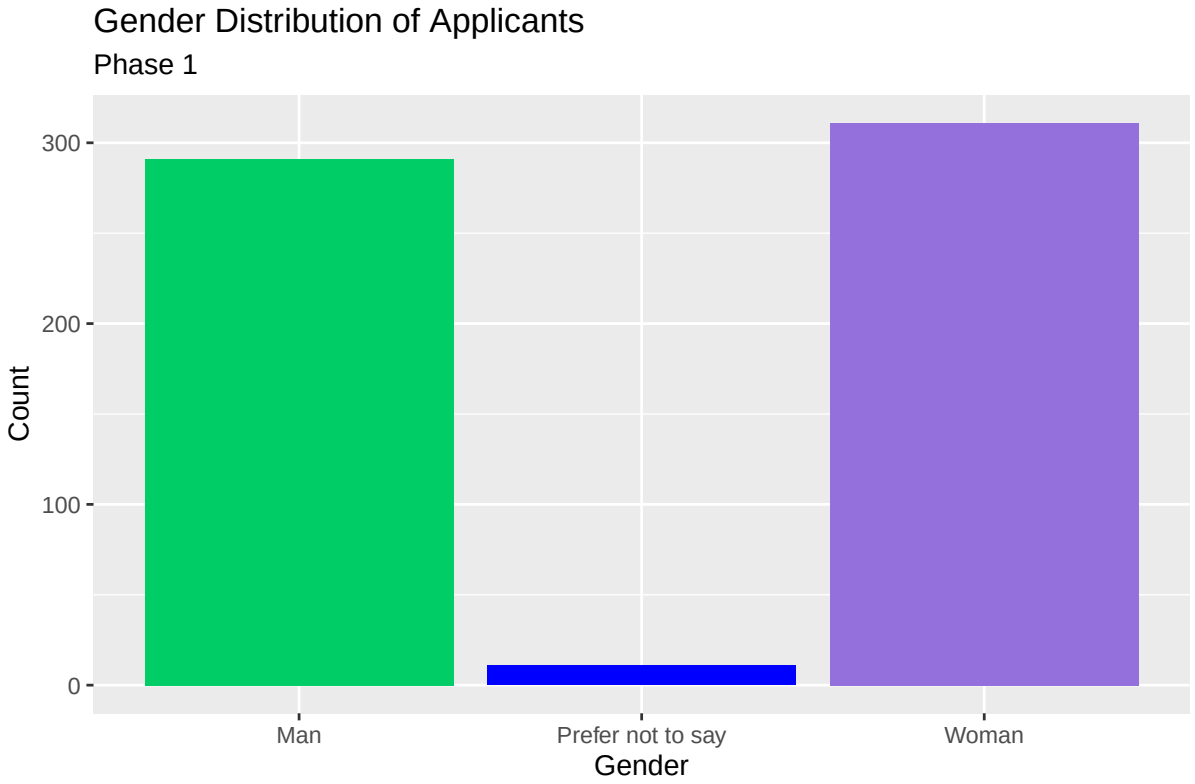

Figure 3.3.2

Within the first phase, there was a nearly even distribution of male and female applicants, as seen in **Figure 3.3.2**, with 291 and 311 respectively, with only 11 non-identifying.

After some analysis, it was found that the necessary requirements for passing this phase were a cv and cover letter, as all applicants that failed to hand in one of these documents were cut going into phase 2. In addition, it was also found that applicants had to have a GPA of at least 2.0 and have scored at least 1 in extracurriculars to pass phase 1. However, meeting these 4 requirements did not guarantee an applicant passed phase 1.

Then, by applying a logistic regression to the data, with a response indicating whether or not an applicant passed phase 1, we were able to inspect the remaining covariates. This model tested; gender, gpa, work experience, and extracurriculars. Through this regression we found that both gpa and work experience were significant factors that benefitted the odds of an applicant's success in this phase. The remaining factors, including gender, appeared to be insignificant.

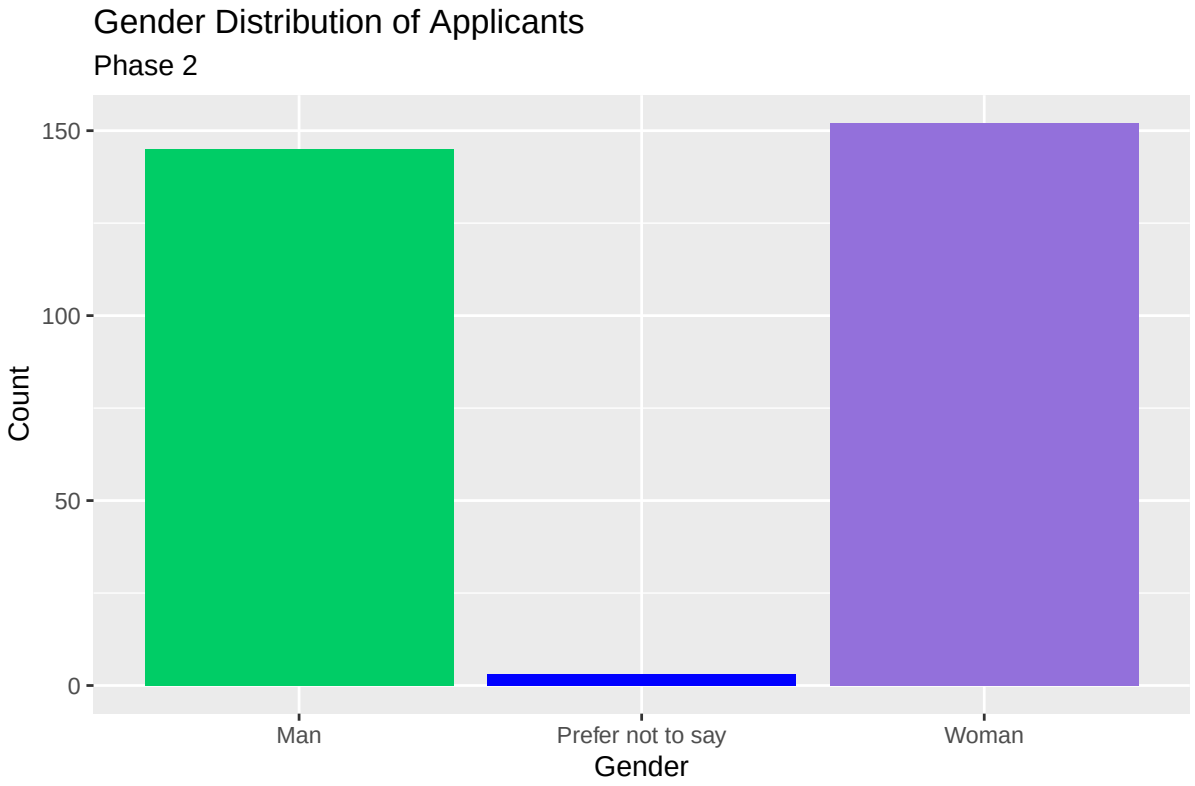

Figure 3.3.3

By the second phase, out of the 300 applicants selected, a similarly nearly even distribution was found with 145 men and 152 women, with 3 non-identifying applicants, displayed in **Figure 3.3.3**. Thus, given the proportions seen and the results from the logistic model, we are fairly comfortable with assuming there is no inherent bias in this section of the hiring pipeline.

That being said, however, some caution should be taken with the extracurriculars question from this phase. While this is difficult to assess from our position, given that this is a significant factor to the model, the company must ensure that the way the AI autogrades this score is fair with its assessment. That is, it must not show any signs of favouring extracurricular activities that might be regarded as "masculine", or conversely, any signs of unfairly scoring activities that it regards as "feminine", as these could end up acting as proxies for gender that affect an applicant's outcome, but might otherwise go unnoticed.

Phase 2 to Phase 3

In phase 2, the AI comes in again to assess a variety of the candidates' performances on multiple tasks. These ratings will be of high importance as they are often able to hide biases behind them quite effectively.

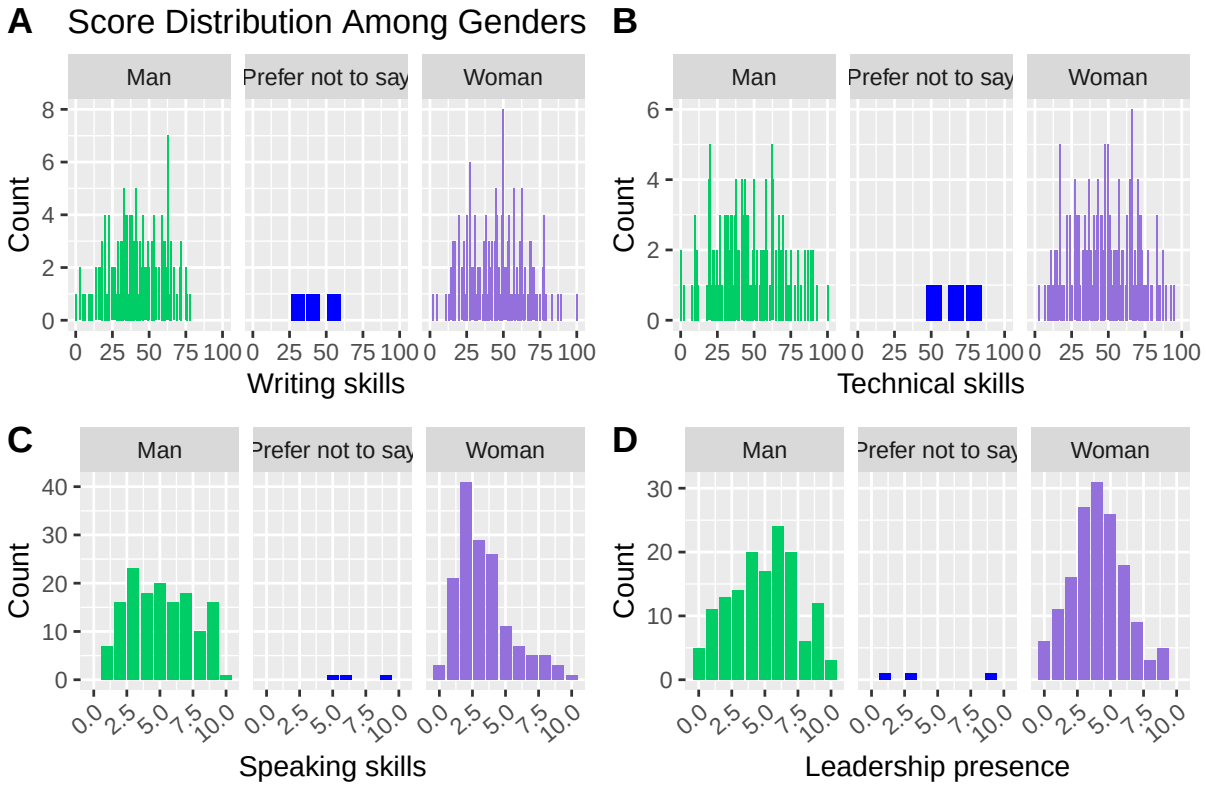

Figure 3.3.4

First, through analyzing **Figure 3.3.4** above, we see that both writing and technical skills (Plots A and B, respectively) have fairly similar score distributions for men and women with averages of 41.8 and 46.2 for writing, and 46.4 and 48.0 for technical skills, respectively. This suggests that the AI is treating these responses relatively fairly. However, when we move to looking at the distributions between of speaking skills and leadership presence (Plots C and D, respectively), a striking difference becomes clear. For leadership presence, women appeared to earn lower scores more often than men did, with average ratings of 4.1 vs. 4.9. Moreover, this difference was extremely noticeable in the speaking skills assessment where women scored an average of 3.3, against the men's average of 5.1. When considering the data that the AI used for these scores, a pattern appears. We find that it tends to give a fair assessment of an applicant's work when their gender is not easily interpretable, as in the writing and technical tasks. But it appears to favour male applicants when gender is easily visible, as it is in the pre-recorded video it analyzes to determine leadership and speaking ability. This is likely the reason behind the difference we see through these 4 scores. Thus, this gives some initial evidence of a possible point of strong bias in the hiring pipeline.

Table 4: Logistic Model Output

	Estimate	Std. Error
(Intercept)	0.000	42.575
genderPrefer not to say	0.000	Inf
genderWoman	0.567	2.060
speaking_skills	2.045	1.183
leadership_presence	2.450	1.227
technical_skills	1.084	1.021
writing_skills	1.097	1.024

To identify the role that these skills play in an applicant's success in the second phase, another logistic regression was fitted to the data, on a response indicating a pass to the next phase. This time, the model included gender, as well as all the AI graded skills from phase 2.

The results from this model, shown in **Table 4** indicated that gender was not a significant factor in determining an applicant's success, but each of the AI graded skills showed strongly positive effects on an applicant's odds of success. We saw that, each unit increase in an individual's speaking skills score and leadership presence, increased the odds that the applicant will pass phase 2 on average by a factor of 2.05 and 2.45 respectively. In addition, the 95% confidence interval for the estimates of speaking skills and leadership presence predicts these values to be between 1.52 and 2.95, and 1.72 and 3.90, respectively. Thus we can safely reject their null hypotheses that these factors' true effect on passing phase 2 is zero. As a result, this further strengthens the claim of possible gender bias against women in the hiring process. Since speaking skills and leadership presence are only graded on a scale from 0 to 10 in comparison to writing and technical skills which are graded out of 100, we can see a greater effect on the odds of success for an increase in 1 point on scores.

While this is a pretty normal finding, when we consider the uneven distribution between men and women in both leadership and speaking skills, it becomes clear that an indirect bias is being created by the AI's autograding that again favours male applicants and benefits their success.

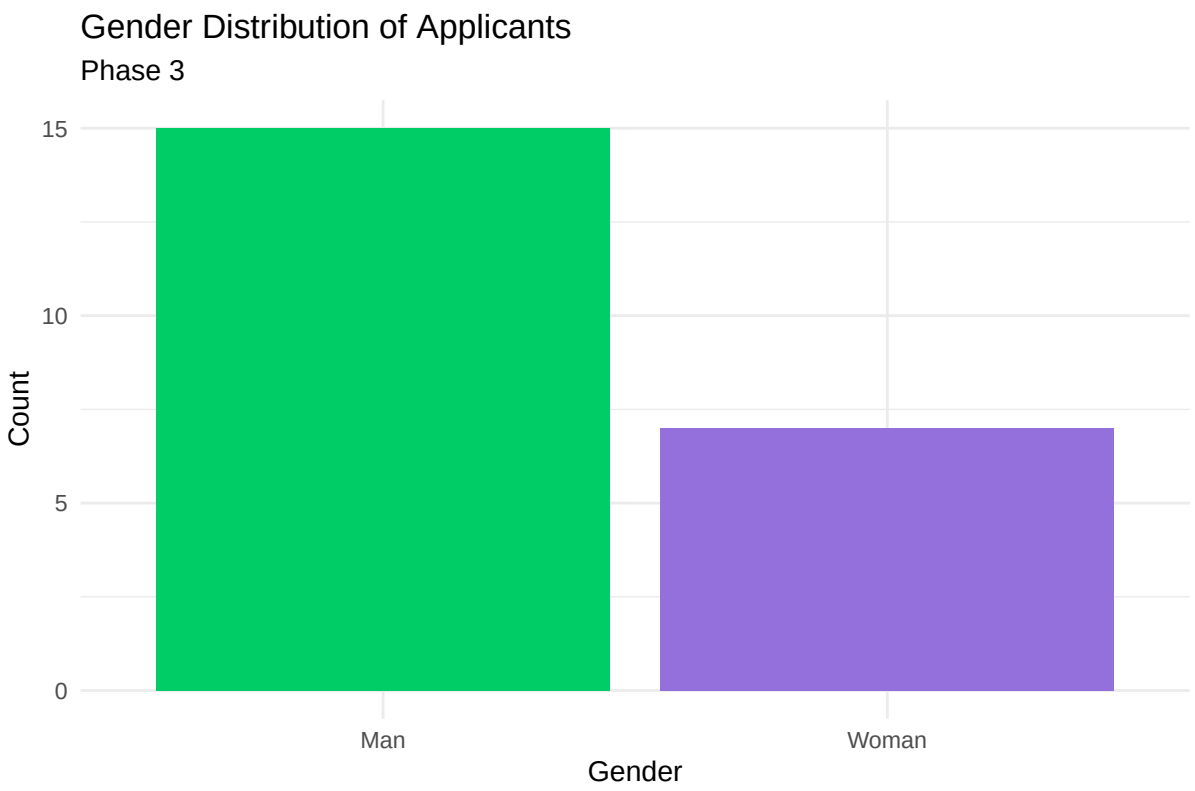

Figure 3.3.5

Moving into phase 3, we find that the once nearly even distribution between men and women has disappeared with 15 men entering the third phase, and only 7 women. At this stage, we can clearly see that, given the strong effects seen in the regression model and the inequality in the AI's ratings, men have gained a significant lead over women and are now seen to dominate the third phase. This is a serious problem as it is a distinct point at which we can detect a strong bias being developed within the hiring pipeline at the company. Thus, we would highly recommend reviewing the AI used to evaluate candidates at this level and make any adjustments to balance these results out.

Phase 3 to Hire

In the final phase, as explained, applicants were rated by two interviewers, and the applicants with the top 5 mean scores from each team were then selected as new hires.

Table 5: Linear Mixed Model Output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.53	8.83	-0.29	0.78
genderWoman	0.59	1.45	0.41	0.69
speaking_skills	2.39	0.57	4.16	0.00
leadership_presence	2.47	0.46	5.39	0.00
technical_skills	0.37	0.04	8.51	0.00
writing_skills	0.38	0.07	5.13	0.00

Table 6: Confidence Intervals

	2.5 %	97.5 %
(Intercept)	-21.26	16.20
genderWoman	-2.49	3.68
speaking_skills	1.17	3.60
leadership_presence	1.50	3.44
technical_skills	0.28	0.46
writing_skills	0.22	0.54

Thus, in order to study the bias here, we took to examining the factors that affect an applicant's mean interview scores. This was done using a linear model on the response of an applicant's mean rating, with covariates for gender, and all of the test scores from the AI. Displayed in **Table 5** and **Table 6** are the estimates and confidence intervals from this model. Similar to phase 3, again, we see that gender itself does not have a significant effect on the mean score an applicant receives, but all testing scores show significant and positive effects on the mean interviewer rating. On average a unit increase in an applicant's speaking skills and leadership presence scores, causes their mean interview scores to increase by 2.39 and 2.47 points respectively. With p-values of 0.0007 and 0.00006, respectively, we are fairly certain that we can reject the null hypotheses here and confirm their significance as they appear to have an effect on mean interview scores.

Thus, we can see that the biases created by the AI are showing up in the interviewer ratings

and taking an effect on the applicants' successes or failures. While we may still be seeing a fair assessment of their "hard" work, the methods for reviewing the more personal skills of speaking and leadership must be investigated and altered accordingly. This again may be the result of a more male dominated hierarchy at the company that has strong tendencies to favour both male employees and applicants.

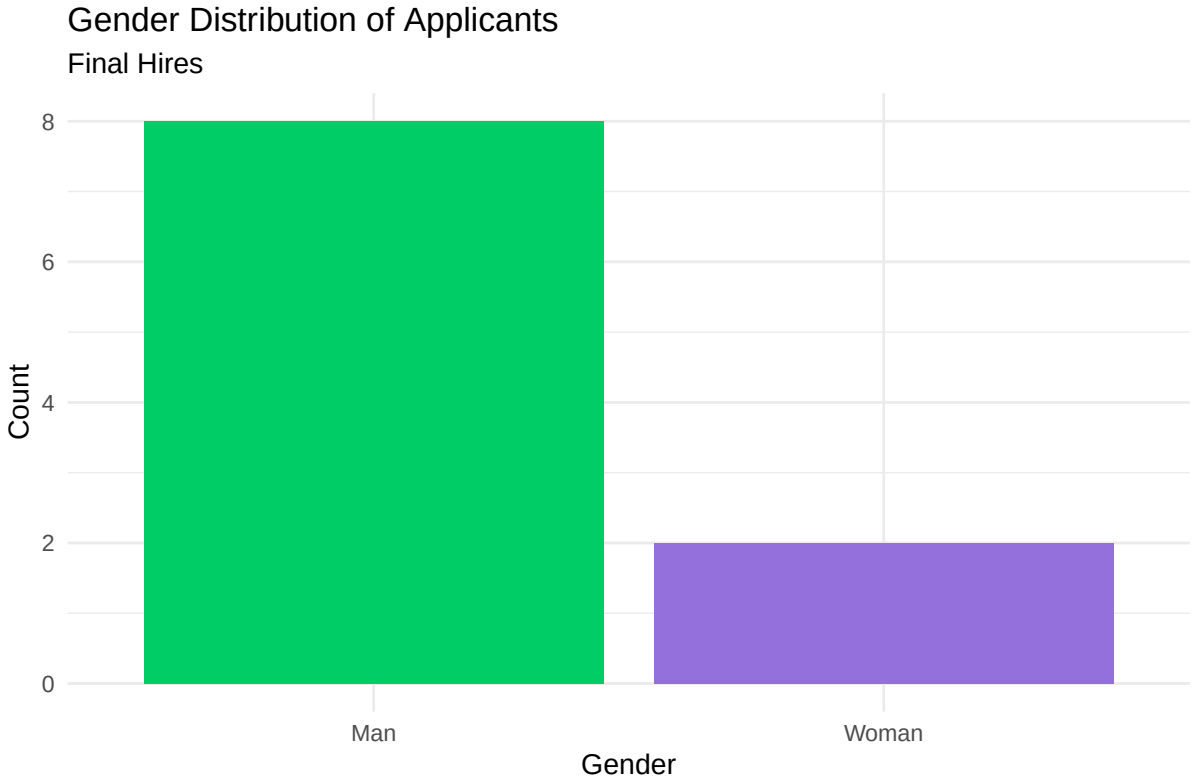

Figure 3.3.6

As we can see, in the final set of hired applicants, the disparity between female and male hires has become even more significant, as out of the 10 applicants selected, only 2 were women while the remaining 8 were all men. Thus we find some serious evidence of a bias against female applicants again in this section of the hiring pipeline that has stemmed from the uneven distribution of test scores from the AI. In addition to having higher mean scores, men also disproportionately received more exceptional scores which comprised the majority of final hires. For applicants that made it to Phase 3, 60% of men scored 7 or better for leadership presence while 42% of women scored 7 or better. For speaking skills, 47% of men scored 7 or better and only 29% of women scored 7 or better in phase 3. As discussed previously men performed much better in these 2 areas, so the source of the bias is most likely how the ai determined scores for these 2 categories. Therefore we believe that the root of bias in the hiring pipeline comes from the AI's autograding in phase 2, which is then passed on to the interviewers at the final stage.

4 Discussion & Conclusions

To state the conclusions from this analysis, they have been broken up to each research question, to be recollected at the end.

4.1 Salary Process

Using the salary data, we found multiple areas that show a bias toward men and thus support the existence of a gender pay gap in the company. Firstly, we found that not only do men make up the largest proportions of employees on the highest paying teams, but across almost all teams in general men appear to take home higher average salaries than women. Moreover, this same effect was seen within the seniority groups at the company wherein men appeared dominant in the highly executive roles. Visualizing the salaries in these groups again showed that men tend to take home greater incomes than women of the same seniority level, providing further evidence of a wage gap. Additionally, when examining the changes in salaries of the employees, we found a trend that suggests men also tend to receive raises much more often than women, adding to this wage gap. Finally, after applying a linear regression to the data, we found that, after controlling for productivity and working time, women on average, earn $2816 less than their male counterparts. Thus the possible existence of a wage gap in the company that favours men appears extremely likely.

4.2 Promotions Process

From the promotion process analysis, we again found some areas that spark concern as they suggest that men may in fact have an easier time gaining promotions in the company. For starters, in the employee reviews themselves, women were the only ones that received a “Needs improvement” rating, while men were the only ones receiving “Exceeds expectations” reviews. This was suggestive of a higher standard that women may be held to in the company that should seriously be reviewed. Furthermore, the distribution of genders within each seniority group showed male dominance in the more executive roles that are most often achieved through promotions. Examining the number of promotions achieved then identified the fact that men earn a much greater number of promotions than women on all counts, suggesting again that men are favoured for promotions in the company. When using a zero-inflated model on the number of promotions we then found that after controlling for working time, being a women decreases the estimated number of promotions by almost 28%. Thus, we are fairly comfortable concluding that there is a bias within the promotions process that makes it easier for men to excel in the company.

4.3 Hiring Process

Through analyzing trends in the data and by applying a regression model to each phase, we were able to gain key insights on how the hiring process was determined and what possible biases were present in this process. Firstly, in phase 1, the initial application phase, it was determined that there was no inherent form of bias present. The nearly even distribution of gender for applicants that made it to phase 2 helped strengthen this claim as gender was not found to be a significant factor. However, in phase 2, it was found that there was bias present in the determination of scores for certain tasks. As previously mentioned, scores for writing and technical tasks were evenly distributed for both men and women alike. However, for speaking and leadership skills, men tested significantly better than women. From these results we can see a clear pattern emerge. We find that for testing where gender is not easily interpretable, the AI tends to give a fair assessment of an applicant's work, as in the writing and technical tasks. But for testing where gender is easily determined, such as in the pre-recorded video that is used to grade speaking skills and leadership presence, the AI system tends to favour male applicants. Factors such as tone of voice, in which men and women ofter differ quite heavily, could have been one of the key factors in determining these scores that would result in a gender bias. Similarly for phase 3, there was once again a decrease in the proportion of women that passed this phase and were hired. Again, speaking skills and leadership presence were significant factors which provided a large source of bias. However, for this phase interview scores were determined by people and not AI. Nevertheless, the interviewers could have still been grading applicants based on the same standards outlined by the company that was used to create the AI system. Qualities such as a lower pitch of voice, which is more common in men, could have been perceived as better speaking skills or leadership presence, as they give off more resonance and thus can be perceived as being louder and more confident.

4.4 Limitations and Future Work

While the data supplied was quite generous as it allowed for many analyses, there were still a few points at which our abilities were limited. Firstly, within the current employees data, both the leadership and productivity ratings generated some concern. We observed some biases in these scores that appear to favour the male employees as they were seen achieving disproportionately greater scores. However, we do not have information regarding which sectors of the company are responsible for these reviews, which limits our ability to provide a specific solution to the problem. Furthermore, the lack of objectively scored variables also reduces the full scope of our models. Including additional factors into the data would have allowed for much more thorough analyses and more generalizable conclusions. This would not only let us comment more confidently on gender biases in the company, but possibly other areas of bias as well. Understandably, however,

providing more detailed information on employees and applicants is a very sensitive topic and should carefully be considered. Furthermore, as we saw some difficulties in our zero inflated Poisson model for the promotional series, we were slightly limited on the conclusions we could generate from it. By possibly having more data that we could use to understand the excess zeros, or taking some alternative statistical methods, we may have been able to produce better results and provide more information on the biases in this process.

The most prominent limitation for the hiring pipeline study is that we are unable to view and analyze the process the AI uses to determine scores for testing in phase 2. Specifically, not knowing which factors the AI utilizes to grade speaking skills and leadership presence when looking at the pre-recorded video applicants submitted. This leaves us to speculate which specific factors are the primary source of bias and prevents us from making a full recommendation on how to limit gender bias in the hiring process.

Based on these limitations and further speculation, there are many possible areas for future work on this topic. The easiest next steps would be to collect more information to include in the analyses that can better explain biases within the company on a wider scale. Moreover, it is likely that the addition of this information will also boost the confidence of the results already generated. Thus, this is a very promising area for future work. Aside from employee and applicant information, additional information on the inner workings of the company would greatly benefit our analyses. Most importantly, gaining more of an understanding of the AI used in the hiring pipeline would allow us to pin down the reasons for the biases we observed. Overall, these improvements would give us a better ability to generate strong conclusions and make more decisive recommendations for actions the company can take to defend against biases.

5 Consultant Information

Consultant Profiles

Timothy Regis. Timothy is a junior consultant with Brownwall Consulting. He specializes in report writing, technical analysis and data organization. Timothy earned his Bachelor of Science, Majoring in Economics and Statistics, from the University of Toronto in 2020.

Kashaun Eghdam. Kashaun is a junior consultant with Brownwall Consulting. He specializes in reproducible analysis and data cleaning. Kashaun earned his Bachelor of Science, Majoring in Economics and Statistics from the University of Toronto in 2020.

Code of Ethical Conduct

Brownwall Consulting is dedicated to ensuring the highest standards of ethical statistical practice and strives to promote this quality in all facets of our company. A few of the guidelines we abide by include:

- We follow all procedures that protect both the rights and interest of individuals involved. Specifically, privacy laws and standards set out by relevant bodies will be respected when collecting and holding information and when publishing findings.
- All intellectual property from fellow statisticians or other sources will be respected and accredited if utilized in our work.
- We will hold ourselves accountable for all work conducted and in the case of a professional assessment or review, all individuals involved will be fully transparent and will not release false information regarding procedures or methods used.