

---

# **New talent acquisition systems and current employees promotion processes analysis**

Report on potential bias at Black Saber Software

Report prepared for Black Saber Software by Core Data  
Consultancy

2021-04-28

## Contents

<b>Executive summary</b>	<b>3</b>
Background . . . . .	3
Key findings . . . . .	3
Strengths and limitations . . . . .	4
<b>Technical report</b>	<b>5</b>
Introduction . . . . .	5
RQ1: Is the AI-automated hiring process biased against any category of gender? . . .	5
RQ2: Are the systems used for promoting employees free from bias? . . . . .	10
RQ3: Is salary calculated according to employees' performance? . . . . .	13
Discussion . . . . .	14
<b>Consultant information</b>	<b>18</b>
Consultant profile . . . . .	18
Code of ethical conduct . . . . .	18
<b>Appendix A</b>	<b>19</b>
<b>Appendix B</b>	<b>19</b>
<b>Appendix C</b>	<b>20</b>

## Executive summary

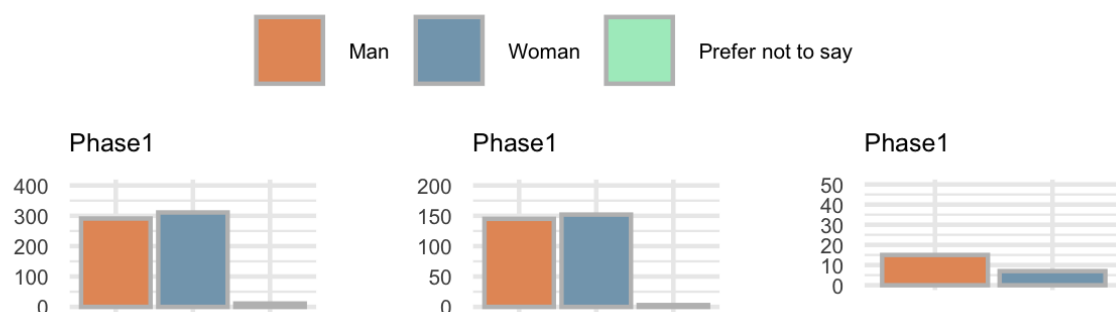
### Background

In 2020, 613 new graduates applied to open positions in Data and Software teams of Toronto-based Black Saber Software. The candidates have gone through three phases in the recruitment process, which were completely assessed by a piloted AI-automated system. This report analyzed concerns raised by Black Saber Software to uncover hidden biases in their AI-automated hiring along with human assessed promoting and remunerating processes. The concerns were formulated in three main questions: (1) Is the AI-automated hiring process biased against any category of gender? (2) Are employees promoted fairly based on their talents and professional achievements? (3) Is salary calculated according to employees' performance?

### Key findings

Regrettably, the analysis exposed a strong unfairness toward female candidates and employees manifested both in the AI-automated systems and human-led assessments.

- Exploratory data analysis revealed that despite of roughly similar skills scores, twice as fewer female candidates reached the third phase of recruitment (Figure 1).



Prepared by Core Data Consultancy  
2021-04-28

**Figure 1:** Gender categories were represented equally in the first two phases with the third phase losing the “Prefer not to say” category and twice as low women than men

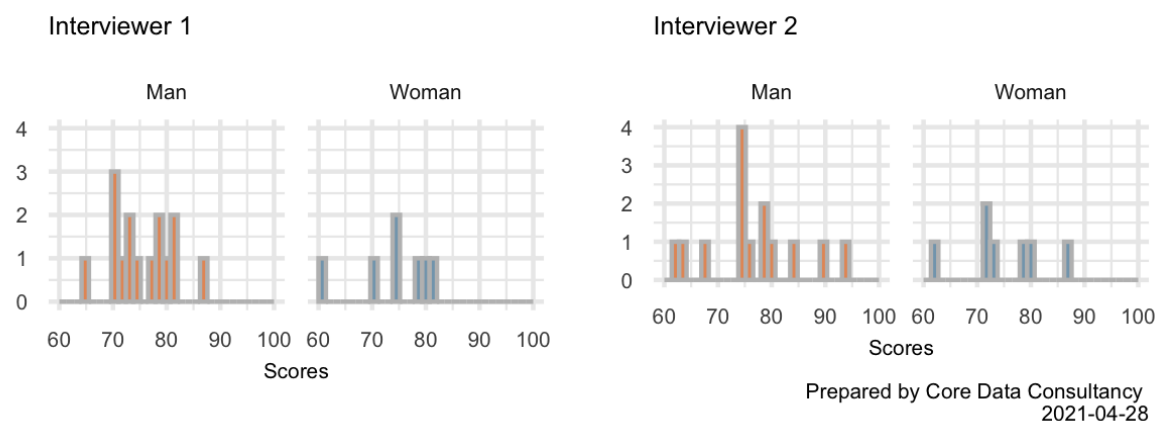
- A statistical model revealed that women lose about 10% in their leadership and speaking skills score due to the means of collecting assessment data - a pre-recorded video was employed, easily revealing candidates' gender association.

- A statistical model predicted that women employees had 60% lower odds of being promoted.
- To explore salary calculation criteria, a test on several models selected employees' gender as a main influential category. Thus, female employees were predicted to receive \$2,642 less than their male co-workers.

## Strengths and limitations

Due to various external reasons and my professional judgement of appropriateness of choices, the report demonstrated the following strengths and limitations.

- The report is completely reproducible with easy adjustment of the models.
- Beyond the scope of the analysis, I have uncovered some unsettling patterns in human interviewers' behavior in the last phase of the hiring. As per Figure 2, the interviewers consistently scored men higher than women.



**Figure 2:** Human interviewers' scoring of men and women candidates

- Limitation of the analysis to four years of the current employees' data might have influenced the overall outcomes of the tests.
- For research question 2, instead of (or in addition to) number of promotions, exploring an increase in salaries might have produced alternative results.

## Technical report

### Introduction

Black Saber Software has piloted an AI service to screen job applicants via timed technical and writing tasks as well as pre-recorded videos to assess effectiveness of candidates' communication and leadership skills. This report analyzed the AI-automated candidates selection process with the main emphasis on uncovering hidden biases against particular categories of gender as that is the only concerning element currently, since People and Talent department had not yet collected data on candidates' and employees' ethnicity. Additionally, Black Saber Software, being aware of recent scandals in media regarding potential bias in remuneration processes in other firms, requested a comprehensive analysis of the company's promotion and remuneration processes to ensure that the current employees are compensated and promoted according to their talent and professional abilities.

### Research questions

To address Black Saber Software's concerns regarding their recruitment and compensation methods, the following research questions were stated.

- Is the AI-automated hiring process biased against any category of gender? Here, I employed four datasets on new graduates who applied in 2020 and explored three recorded gender categories - "Man", "Woman", and "Prefer not to say".
- Are employees promoted fairly based on their talents and professional achievements? Analysis of this question was defined by the data on current employees collected quarterly from 2013 to 2020.
- Is salary calculated according to employees' performance? The same current employees' data as in the previous research question was used. Such metrics as productivity and leadership skills among others will be evaluated to find an evidence pro or against gender bias.

### **RQ1: Is the AI-automated hiring process biased against any category of gender?**

To answer the topic question, I found Phase 2 data most helpful as it contained the AI-rated metrics such as technical, writing, speaking, and leadership skills. Since the AI assessed the technical skills similar to the writing based on a timed task and measured on the same scale from 0 to 100, I combined the skills in one column to use as a response variable. Similarly, I have grouped the speaking and leadership skills as they were estimated via pre-recorded videos

on identical scale from 0 to 10. Figure 3 illustrates that for technical, writing, and leadership tasks most of the candidates were given a score between 25th to 75th percentiles (the variables follow a normal distribution); the scores for the speaking skills were slightly right-skewed.



**Figure 3:** For technical, writing, and leadership tasks most of the candidates were given a score between 25th to 75th percentiles (the variables follow a normal distribution); the scores for the speaking skills were slightly right-skewed.

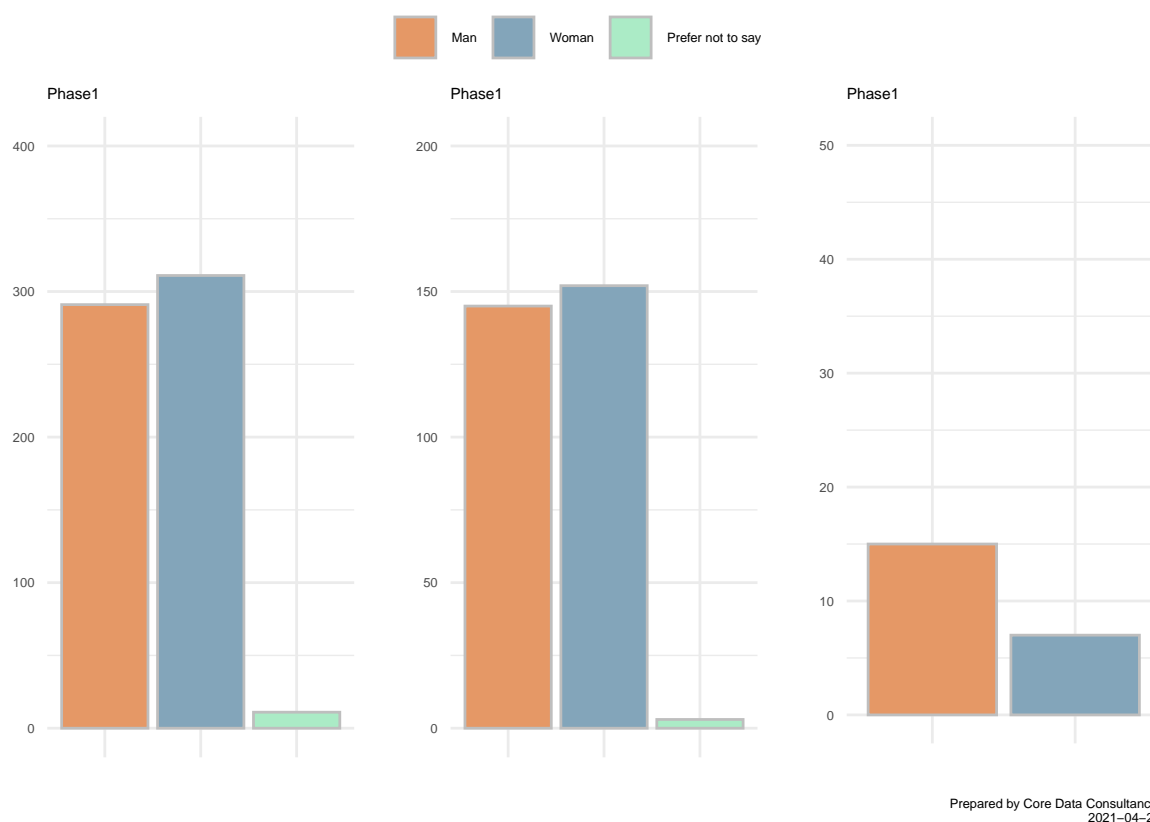
On average, women were ahead of men in technical and writing tasks, and behind men in speaking and leadership skills; those who preferred to not disclose their gender scored higher or at par across categories (Table 1). I will get back to the results reflected in the Table 1 further in the analysis.

Before diving into modeling, the following exploratory analysis paints some valuable patterns in gender distribution across various categories such as hiring teams, candidates' extracurricular activities, and their work experience and GPA. Please note that as the number of the candidates reduce drastically from phase to phase, the scales of the following several visuals were not necessarily equal to emphasize the comparison only.

**Table 1:** Average values for main parameteres of interest

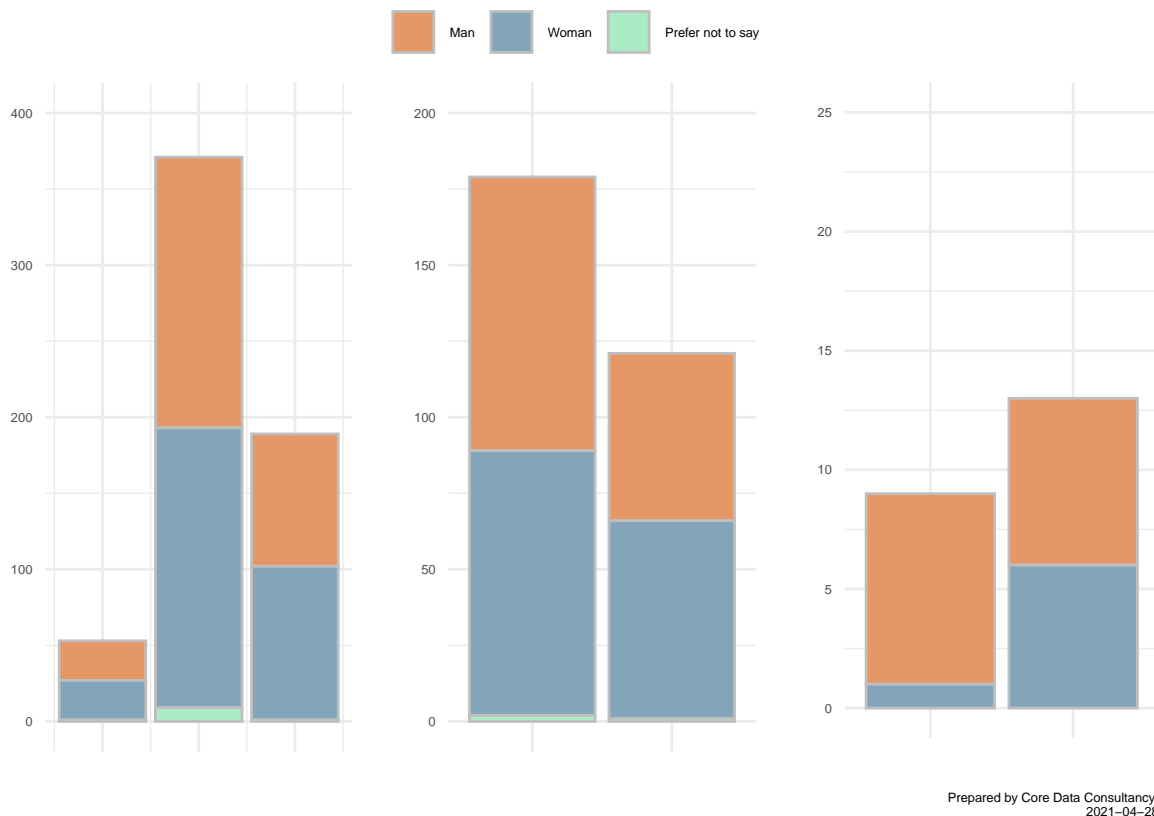
Gender	GPA	WorkExp	Technical	Writing	Speaking	Leadership
Man	3.084828	1.137931	46.37931	41.77931	5.075862	4.924138
Prefer not to say	2.666667	1.000000	66.00000	42.33333	6.666667	4.333333
Woman	3.138816	1.171053	47.98026	46.19079	3.309210	4.065790

Figure 4 demonstrates that all gender categories were represented equally with strong skewness toward men candidates and disappearing of the undisclosed gender category in the phase 3.



**Figure 4:** All gender categories were represented equally with strong skewness toward men candidates and unrepresented undisclosed gender category in phase 3

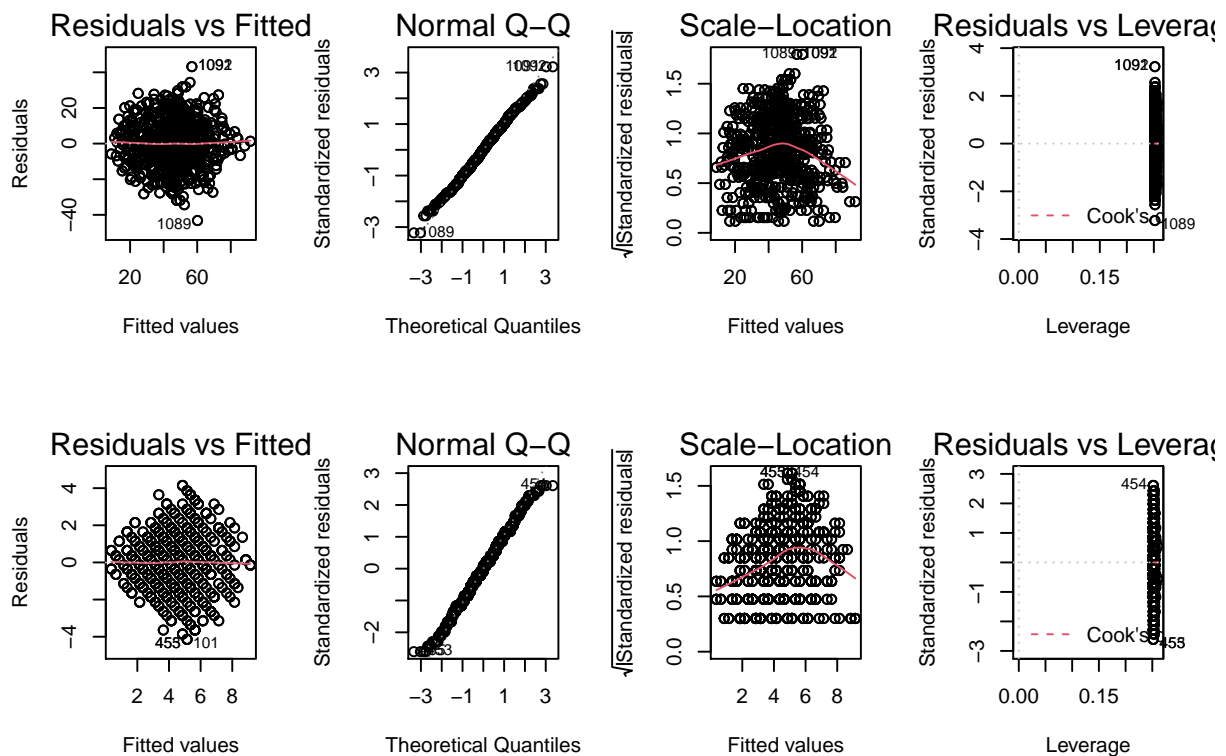
Candidates' extracurricular activities were coded as 0 for no extracurriculars, 1 for “some relevant and/or skills building extracurriculars”, and 2 indicating “several high relevance and/or skills building extracurriculars”. In phase 3, there were considerably more men in the lower extracurricular category (Figure 5).



**Figure 5:** In phase 3, there were considerably more men in the lower extracurricular category.

Since it was clear from the preceding exploratory analysis that in spite of similar ranking, Phase 3 gender distribution was skewed toward “Man” category (Figure 4). To further investigate the trend and answer the first research question - Is the AI-automated hiring process biased against any category of gender? - my first choice was a linear model as the outcome of interest was candidates’ skills scores measured on a continuous scale. Since I introduced two response variables - combined technical and writing as one, and leadership and speaking as second - I fitted two regressions and reported two separate results (referred to as Regression 1 and Regression 2). The predictors for both regressions were applicants’ gender, their GPA, extracurricular activities, and work experience. After comparing nested models to find the best fitting one, the predictors differed slightly for two models. Further, Figure 6 demonstrates that all linear regression assumptions were met, except for independence assumption as there were repeated measures for each applicant. Considering non-independence between groups and the repeated measures, I narrowed down my model choice to a linear mixed model with random intercepts estimated for each applicant.





**Figure 6:** Linear regression assumptions check for RQ1.

### Results of Regression 1:

The final best model that investigated whether the technical and writing skills (measured from 0 to 100) of the candidates were assessed fairly used applicants' gender and GPA score as predictors. The skills were given a score of 14 points more for persons of unknown gender, and two points more for females compared to males with a baseline score of 13 points. Each point increase in the GPA score would have led to a 10-point increase in the overall skills score (Table 2).

**Table 2:** Regression 1 results: The skills were given a score of 14 points more for persons of unknown gender, and 2 points more for females compared to males with a baseline score of 13 points.

	Estimate	95%CI
Baseline	13.00	(3.07, 21.97)
Undisclosed gender	14.00	(-1.86, 30.59)
Woman	2.00	(-0.77, 5.68)
GPA	10.00	(7.26, 13.20)

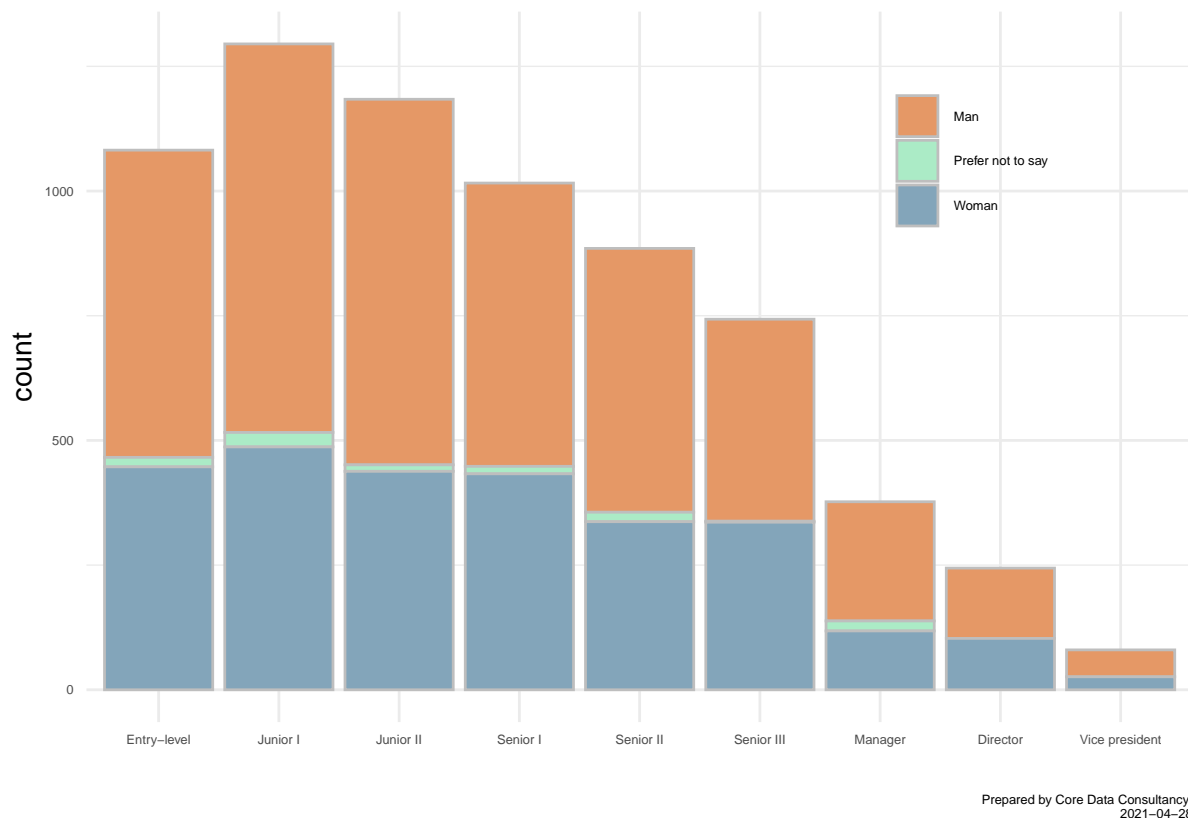
### Results of Regression 2:

The final best model that investigated whether leadership and speaking skills (measured from 0 to 10) of the candidates were assessed fairly, used applicants' gender, GPA score, and their extracurricular activities as predictors. In this model, applicants of undisclosed gender would receive about half a point more, whereas women would lose one point, which equals to 10% considering the scale. Each point increase in the GPA score would lead to about 1% increase in the overall skills score; the highly relevant extracurriculars would get 0.5 point more than less relevant ones (Appendix A).

Possibly, the women candidates possessed weaker speaking and leadership skills, which showed in a recorded video. However, there was a chance that during the evaluation the AI exhibited a historic pre-existing bias against females being not presented as powerful leaders and influential speakers. The results of the Regression 1 potentially could serve as proof to the previous point as there was a very low to zero chance of recognizing gender through a technical or writing task.

### RQ2: Are the systems used for promoting employees free from bias?

Although the current employees' data provided by the People and Talent department contained records starting from 2013, I have decided to include four years of data from 2017 to 2020 as the earlier data was very sparse. To address the second research question, I have created two additional variables to reflect increase in salaries and number of times employees have been promoted over the years. Figure 7 below shows that although there were more men than women in the company, gender was distributed evenly across the roles.



**Figure 7:** Although there were more men than women in the company, gender was distributed evenly across the roles

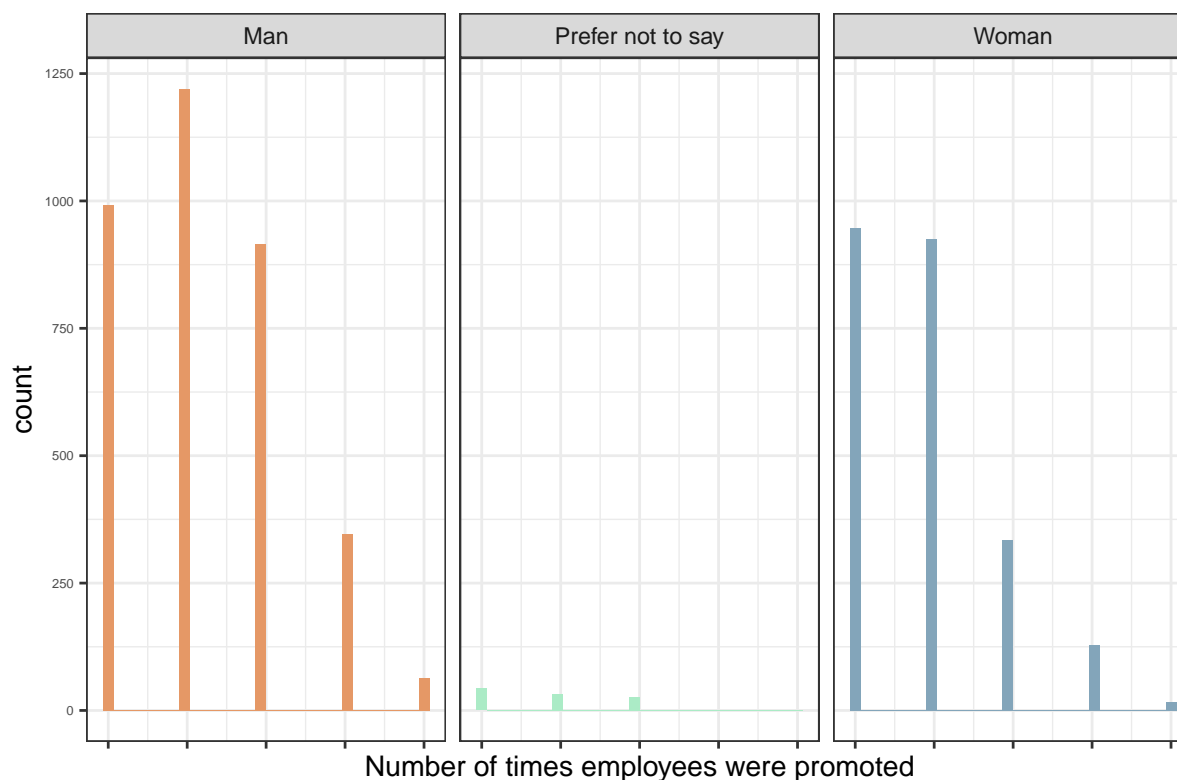
Since the variable of interest was a number of times a person has been promoted over the years (a count variable) with a right-skewed distribution and a repeated measure of employees needed to be accounted for, a generalized linear mixed model with a Poisson distribution was a good fit. Additionally, many employees were not promoted yet because they joined the company very recently, suggesting that zero-inflated Poisson (ZIP) regression would be a better fit. I ran both GLMM and ZIP to test which model would be more improved and found out that the results of the ZIP were not statistically significant (Table 3). I proceeded with the GLMM, establishing that the outcome variable was not overdispersed, which satisfied a key Poisson assumption (Table 4). Further, the second assumption - Poisson response - was proved to be satisfactory as the response can be reasonably modeled with a Poisson distribution when grouped by a key explanatory variable: gender (Figure 8).

**Table 3:** Results for ZIP: the predictors are not statistically significant

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.2055197	0.0151736	13.544600	0.0000000
genderPrefer not to say	-0.3760782	0.1089257	-3.452613	0.0005552
genderWoman	-0.3454414	0.0268231	-12.878513	0.0000000
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-12.152041	35.60905	-0.3412628	0.7329058
genderPrefer not to say	1.073162	180.03064	0.0059610	0.9952438
genderWoman	-1.779465	114.80026	-0.0155005	0.9876329

**Table 4:** Established that the outcome variable was not overdispersed: mean = variance

gender	mean	sd	var	n
Man	1.2281595	1.0205573	1.0415371	3537
Prefer not to say	0.8431373	0.8174471	0.6682198	102
Woman	0.8694173	0.8984410	0.8071962	2351



**Figure 8:** The second assumption - Poisson response - proved to be satisfactory: grouping the response by a key explanatory variable - gender

### Results of GLMM regression:

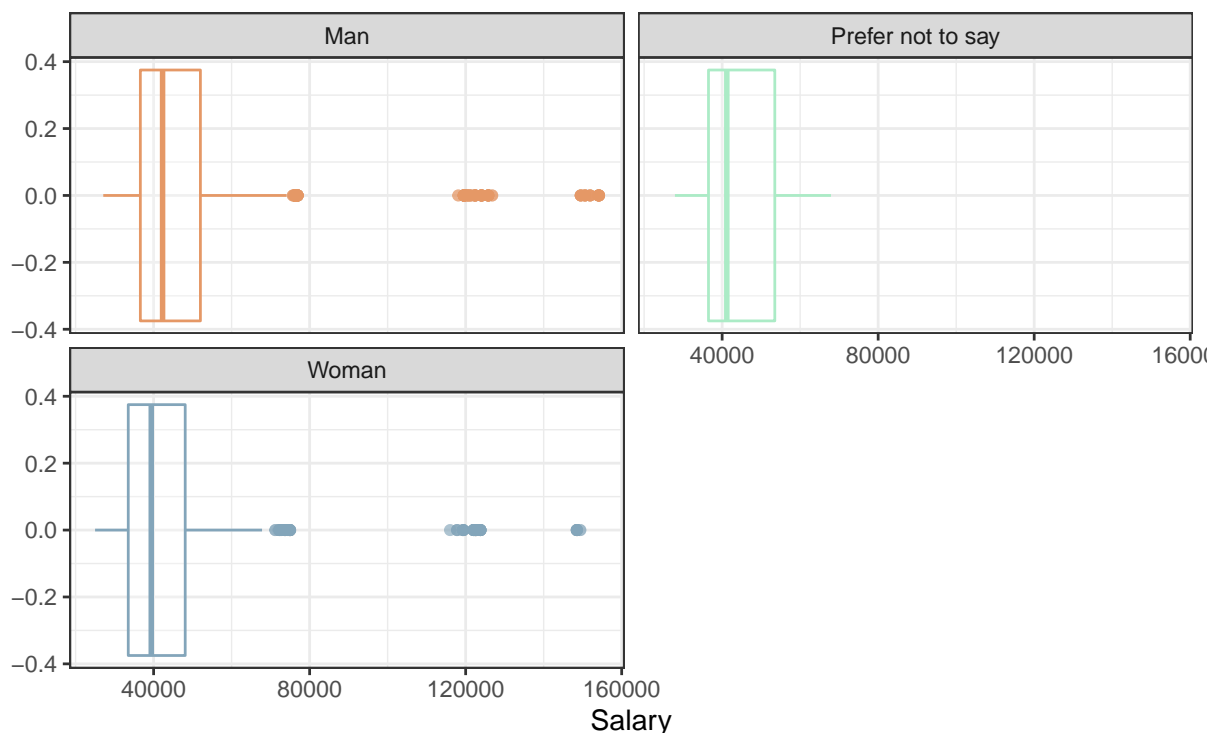
The GLMM with a Poisson distribution suggested the following results:

- for “Prefer not to say” gender:  $\log(\lambda) = -1.27 - 0.45 = -1.72$ , and
- for women:  $\log(\lambda) = -1.27 - 0.9 = -1.17$ .

The final best model that investigated whether employees were promoted fairly used employees’ gender as a predictor. In this model, the odds of being promoted were about 0.3 to 1. Employees of undisclosed gender had 37% lower odds of getting promoted, and women had 60% lower odds of being promoted (Appendix B).

### RQ3: Is salary calculated according to employees’ performance?

To test this question, a linear mixed model proved to be the best fit, since a response - salary - is a continuous variable (Figure 9), following a normal distribution. The model accounts for several observations for each employee and salary differences according to job roles, too.



Prepared by Core Data Consultancy  
2021-04-28

**Figure 9:** Salary distribution across gender categories

### **Results of the LMM Regression:**

The final best model that investigated whether salary was calculated based on employees' performance was chosen by running a test on nested models and used employees' gender as a predictor. The baseline salary was \$68,422. The employees who preferred to not disclose their gender had 726 dollars decrease in the salary. The model predicted even steeper salary decrease of 2,642 dollars for female employees. All these results were reflected in the Appendix C.

```
## Computing profile confidence intervals ...
```

### **Discussion**

This report analyzed three questions raised by Black Saber Software to uncover hidden biases in their hiring, promoting, and remunerating processes. Regrettably, the analysis exposed a strong unfairness toward female candidates and employees manifested both in the AI-automated systems and human-led assessments.

I discovered some evidence that the new AI-automated candidates' selection system might exhibit bias against women when assessing the speaking and leadership skills via pre-recorded videos. This did not come as a big surprise as voice and image recognition systems have been historically demonstrating extreme biases as the underlying databases were created as such. This finding was strengthened by a fact that women's scores on technical and writing tasks did not differ from men's as the tasks were performed without the use of candidates' images or voice.

In the second question of potential biases in the systems used for promoting employees, I have explored a number of times employees got promoted over the four chosen years from 2017 to 2020. These systems also proved to be biased against non-males with 60% lower odds of promotion for women and 37% lower odds for other employees of undisclosed gender.

Finally, while testing several models to detect whether the employees were compensated proportionally to their professional achievements, none of the models picked up on indicators of concern such as employees' productivity, their leadership qualities, or duration of their service in the company. Instead, the successful model only accounted for gender of the employees, proving once again that, unfortunately, salaries were not calculated fairly.

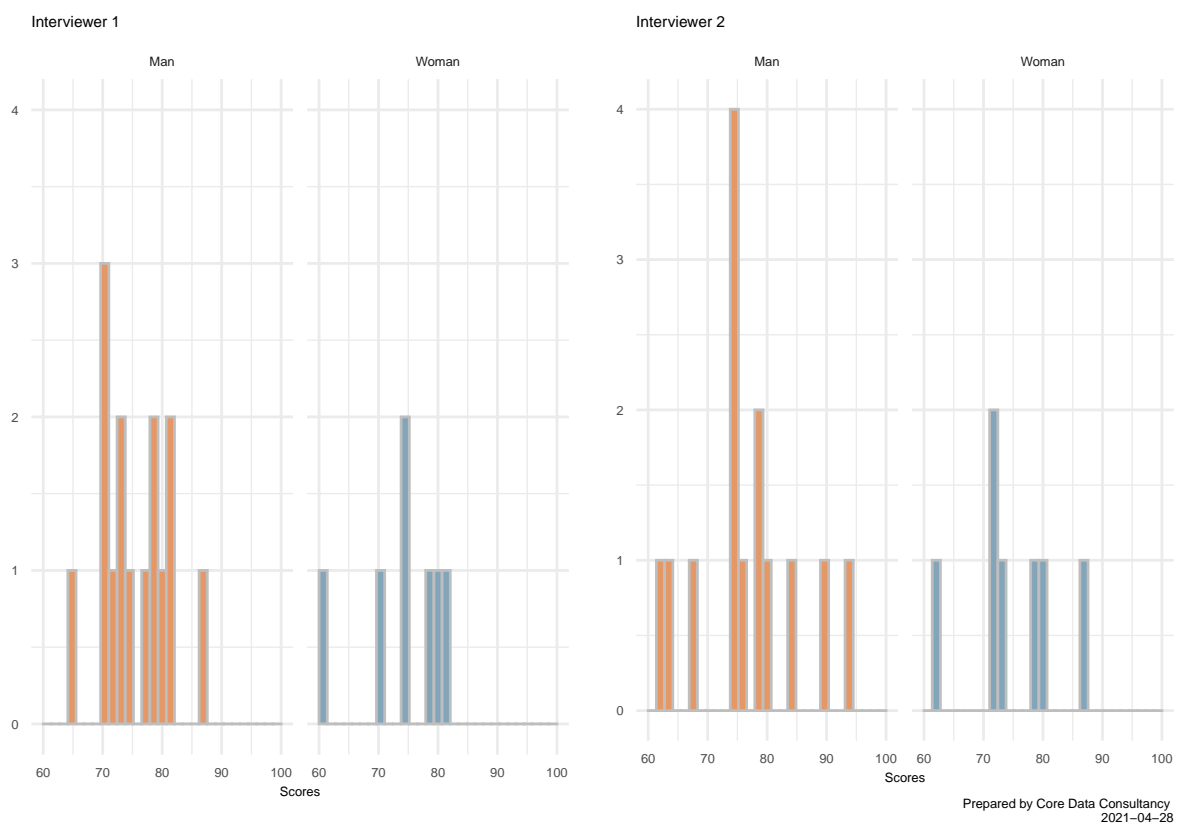
### **Strengths and limitations**

One of the report's strengths was its complete reproducibility and ease of models' adjustment. For instance, whenever there will be an ethnicity metric collected for job applicants and current employees, it can be added in the analysis for further review. In addition, I have discovered

**Table 5:** Interviewer 1 and 2: minimum and maximum scores for male and female candidates

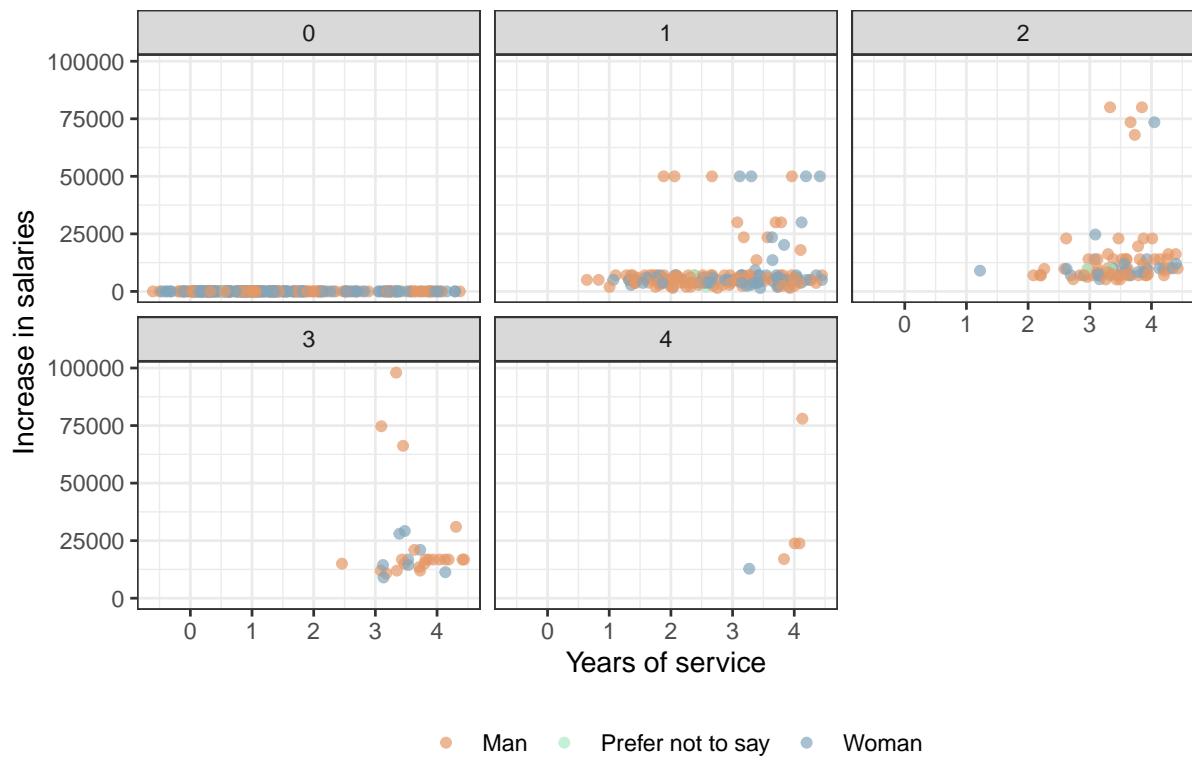
Gender	Min1	Max1	Min2	Max2
Man	65	87	58	94
Woman	60	81	62	87

some patterns in the human interviewers in the last phase of the hiring, which I did not include in the main analysis as this was not explicitly requested. Figure 10 and Table 5 demonstrate that both interviewers tend to score men higher than women. Here, further investigation would be required.

**Figure 10:** Both Interviewer 1 (left) and Interviewer 2 (right) tend to score men higher than women.

I limited the analysis to four years of the data on current employees as 2013 to 2016 had inadequate amount of data. My choice was also supported by the exploration of salary promotions in the original and reduced data. Figures 11 and 12 demonstrated clearly that men were promoted considerably more frequently in both cases. However, this limitation might have influenced the

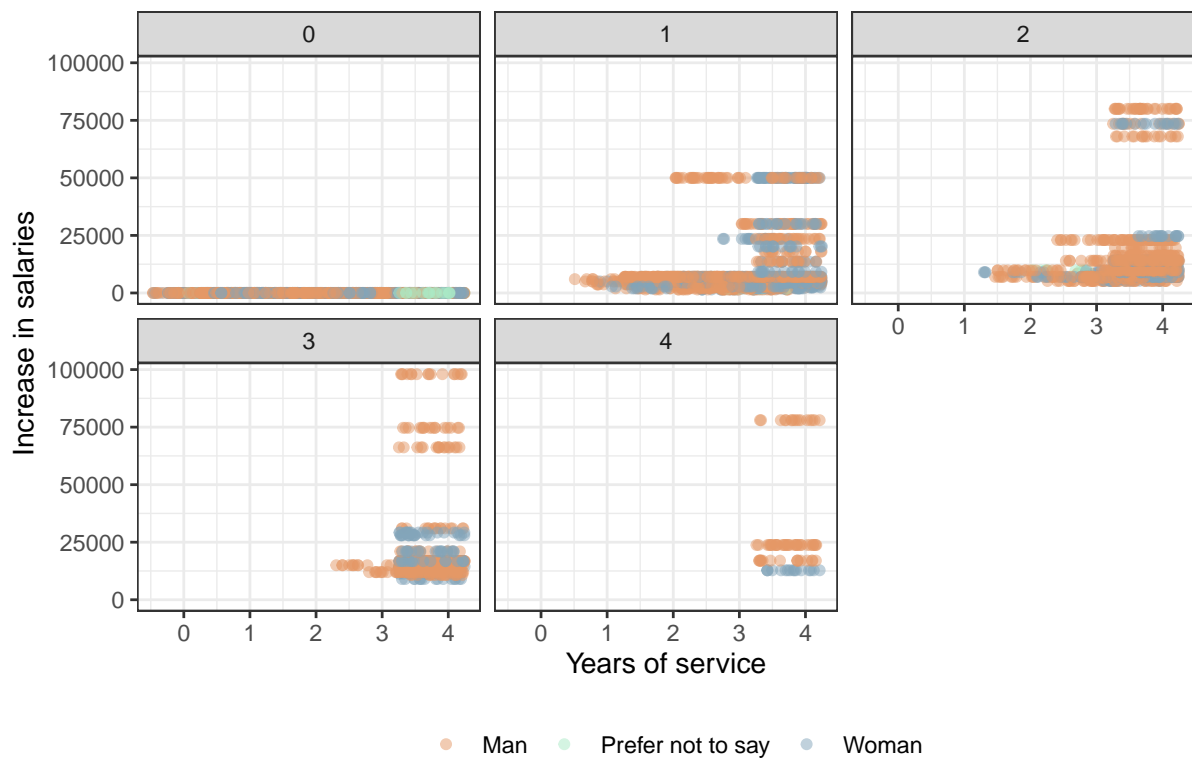
results in an unpredicted way.



Prepared by Core Data Consultancy  
2021-04-28

**Figure 11:** Increase in salaries by years of service and number of promotions. Original dataset containing data from 2013 to 2020: Men were promoted more frequently





Prepared by Core Data Consultancy  
2021-04-28

**Figure 12:** Increase in salaries by years of service and number of promotions. In last four years the situation did not improve: still more men were promoted.

In addition, while tackling the research question 2 - Are the systems used for promoting employees free from bias? - I have not explored an increase in salaries as an outcome of interest. This parameter might have behaved differently than the number of times employees got promoted.

## **Consultant information**

### **Consultant profile**

**Asel Kushkeyeva.** Asel has been in an independent consultancy since 2021. Her firm, Core Data Consultancy, has been recognized among leading Canadian consultancies. In 2022, Asel has won a prestigious The Canadian Journal of Statistics Award. She specializes in reproducible analysis, statistical communication, and data visualization. Asel earned her Master of Information, Specialist in Human-Centered Data Science, from the University of Toronto in 2021.

### **Code of ethical conduct**

#### **1. Professional Integrity and Accountability**

At Core Data Consultancy, professional conduct is the priority in all aspects of our work. We take pride in a high level of accountability at all times, striving to always accept our mistakes and ensuring minimized negative consequences for all stakeholders.

#### **2. Integrity of Data and Methods**

Continuously providing professional services, Core Data Consultancy strives for reproducibility of our work to further encourage development in science and data industry. In order to produce quality reports, we ensure data and statistical methods integrity.

#### **3. Responsibilities to Science, Public, and Client**

Being a mature player in the field of independent statistical analysis, Core Data Consultancy holds itself responsible to science, public, and clients by adhering to science and industry ethical standards, considering public interests, and going above and beyond to keep clients' information secure at all times.

## Appendix A

`\begin{table}[H]`

`\caption{ Appendix A. Regression 2 results: Applicants of undisclosed gender would receive about half a point more, whereas women would lose about 10% in score.}`

	Estimate	95%CI
Baseline	4.00	(3.28, 5.42)
Undisclosed gender	1.00	(-1.25, 2.42)
Woman	-1.00	(-1.71, -0.98)
GPA	0.00	(-0.20, 0.48)
Extracurricular	1.00	(0.20, 0.95)

`\end{table}`

## Appendix B

`\begin{table}[H]`

`\caption{Appendix B. GLMM with a Poisson distribution results: Employees of undisclosed gender had 37% lower odds of getting promoted, and women had 60% lower odds of being promoted.}`

	Estimate	95%CI
Baseline odds	0.28	(0.21, 0.37)
Undisclosed gender	0.63	(0.13, 2.94)
Woman	0.40	(0.26, 0.60)

`\end{table}`

## Appendix C

**Table 6:** Results of the salary model: The employees who preferred to not disclose their gender had 726 dollars decrease in the salary; predicted even steeper salary decrease of 2,642 dollars for female employees

	Estimate	95%CI
Baseline	68422.00	(39142.30, 97701.62)
Undisclosed gender	-726.00	(-2931.24, 1479.90)
Woman	-2642.00	(-3210.73, -2072.86)