Bias Analysis in Human Resource System

-for Blacksaber Software in 2021

Report prepared for Black Saber Software by Data Over Flow

2021-04-21

Contents

Executive summary

We (Data Over Flow Co.Ltd.) have examined the structure of human resource system of the company (the Black Saber Software) by analyzing data on the company's hiring, promotion and salary process and found there to be no bias.

The key findings are:

- Even though there exists difference between salaries of male and female employees, the difference is not the evidence of bias because productivity acts as a confounder variable between gender and salary;
- Among the 3 phases of recruitments, the chance of success is dependent on different variables such as gpa and work experience, gender is not a significant predicting variable for the chance of success at any of these phases;
- The chances of being promoted is higher for male employees than for female employees, however (...)

In our opinion, the system is fair during each of the three process, in accordance with Ontario's Human Rights Code and Black Saber's policies. Specifically, neither hiring nor promotion process shows sign of gender/racial discrimination; the individual salary level is fairly evaluated based non-personal and work-related parameters only.

Technical report

Introduction

We assess the existence of bias from 3 aspects, salary levels, hiring decisions and promotion. We use a combination of model buliding and data visualization to prove our conclusion that there is no bias on personal traits in the company's hiring and enumeration processes. In the following report, "the company" refers to the BlackSaber Software and "we" refers to the analysts from Data Over Flow. With no special notes, the assumed level of significance in this report is $\alpha = 0.05$.

Research questions

- Salary
- Hiring
- Promotion

Does there exist bias in current employee enumeration?

In order to analyze the existence of bias in the current employee enumeration, we fitted a linear mixed model as well as two general linear model on the dataset containing all current employees at the BlackSaber Software. The dataset contains 6909 observations of 9 variables, including all 608 employees of the BlackSaber Software. Each observation records the information of a specific employee at a specific timespot. Based on data visualization, the difference inbetween different level of juniors and seniors are not significant, thus we combine "Entry-level", "Junior I", "Junior II" and "Junior III" into "Junior" and combine "senior I", "senior II", "senior III" into "senior".

We consider salary as the key of measuing employee enumaration and use it as our response varible. We want to asses whether an employee's salary is affects by his/hers gender in order to detect the existence of bias. On top of that, we include other non-personal variables such as team information, financial quater of the year and seniority of level which are ket factors of salary level given common senses. To incorporate both the fixed effetcs and the random effects of variables on the employee enumaration, we choose to use the linear mixed model:

$$y = X\beta + Zu + \epsilon \quad (1)$$

Linear mixed model considers the response variable y as a vector of observations with $E(y) = X\beta$. β is an unknown vector for the fixes effects and u is an unknown vector for the random effetcs. In our initial model, we consider gender, seniority and the financial quarter as fixed effects; we treat team, leadership for level and productivity score as random effects. This allows the model to account for both the variability within random effect groups as well as between those group, which is crucial in our case because the data covers information across various time intervals.

Due to the high correlation between random effect variables, we performed likelihood ratio tests on various reduced models and dropped the leadership for level as a predictor. The final model gives the following results:

Table 1: Statistics Summary for Reduced Salary Model

Variable	$\hat{\beta}$	Std. Error	p-value
intercept	116904.03	709.18	0.99
women	-2928.38	426.88	$6.06e^{-9}$
junior	-79703.19	342.56	$8.91e^{-7}$
manager	-50342.84	311.51	0.00
senior	-69910.72	323.8	0.01
vice-president	30024.17	498.25	$1.25e^{-12}$
quarter	74.31	28.28	0.01

Under this model, we predict a positive correlation between seniority level and salary. There is a difference between average salary of males and females. Furthermore, we fitted three general linear models to check the correlation between gender and productivity:

$$y_{productivity} = X\beta_{gender} + \epsilon \tag{2}$$

$$y_{salary} = X\beta_{gender} + \epsilon \tag{3}$$

$$y_{salary} = X\beta_{productivity} + \epsilon \tag{4}$$

According to 2, the average productivity of female employees is 1.46 times higher of male employees holding all other variables constant(p-value = 0.0001). 4 indicates a negative correlation between productivity score and salary level. Together, this means productivity is a confounding variable between gender and productivity.

Does there exist bias in the hiring process?

We assess the bias in the company's hiring process using a hiring dataset. The hiring process contains three phases, where at each phase, different variables are examined by the human resourses department to determine the successfull applicants. Therefore we derive 3 different data sets from the original data and formed data of new applicants at each phase. The variable of interest is whether the candidate made it to the next phase, which is represented by a binary variable. For this reason, we use general linear models with family of binomials to inference the hiring results. The model predicts p in **??** where p is the probability of event A that we are interested in, $\beta_0$is the intercept, $x_1 ... x_K$ are our variables of interest and $\beta_1 ... \beta_k$ are parameters for each of these variables.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{5}$$

At phase 1, the predicting variables include gpa, whether candidate has extracurricular experience, whether candicate submitted a cover letter and whether candidate has working experience. The results are summarized:

Table 2: Statistics Summary for Phase 1 Hiring Model

Variable	$\hat{\beta}$	Std. Error	p-value
intercept	-25.16	648.71	0.96
prefer not to say	0.16	0.85	0.84
women	-0.05	0.22	0.78
gpa	2.09	0.23	$< 2e^{-16}$
extra-curricular	0.29	0.21	0.17
cover letter	-18.68	648.7	0.97
work experience	0.76	0.27	0.01

According to $\alpha = 0.05$, only gpa and work experience affects the result of phase 1, where both variables have a positive relationship with the chance of passing phase 1. Note that gender is not a significant predictor in this model.

At phase 2, the predicting variables include the team which candidate applied for, whether candidate has extracurricular experience, whether candidate has working experience, and three scores representing writing, speaking and technical skills. The results are summarized:

Table 3: Statistics Summary for Phase 2 Hiring Model

Variable	$\hat{\beta}$	Std. Error	p-value
intercept	-24.15	4.79	$4.77e^{-7}$
prefer not to say	1974.74	0.85	0.99
women	-0.63	0.79	0.42
team	1.4	0.76	0.06
extra-curricular	0.71	0.21	0.37
work experience	0.73	0.27	0.88
technical skills	0.02	0.27	$7.06e^{-5}$
writing skills	0.1	0.02	$9.93e^{-5}$
speaking skills	0.9	0.22	$9.13e^{-6}$
leadership presence	1.00	0.21	$3.73e^{-5}$

According to $\alpha = 0.05$,all the skills score and leadership presence affects the result of phase 1, work experience is no longer a predictor for the chance of success at phase 2.

Phase3 and final TBD

To conclude, each phase of recruitment focuses on difference traits of a candidates, and there is no significant relationship between gender and the rate of success. Besides, there is no significant priority out of the three skills recorded.

Is there gender bias in the promotion process?

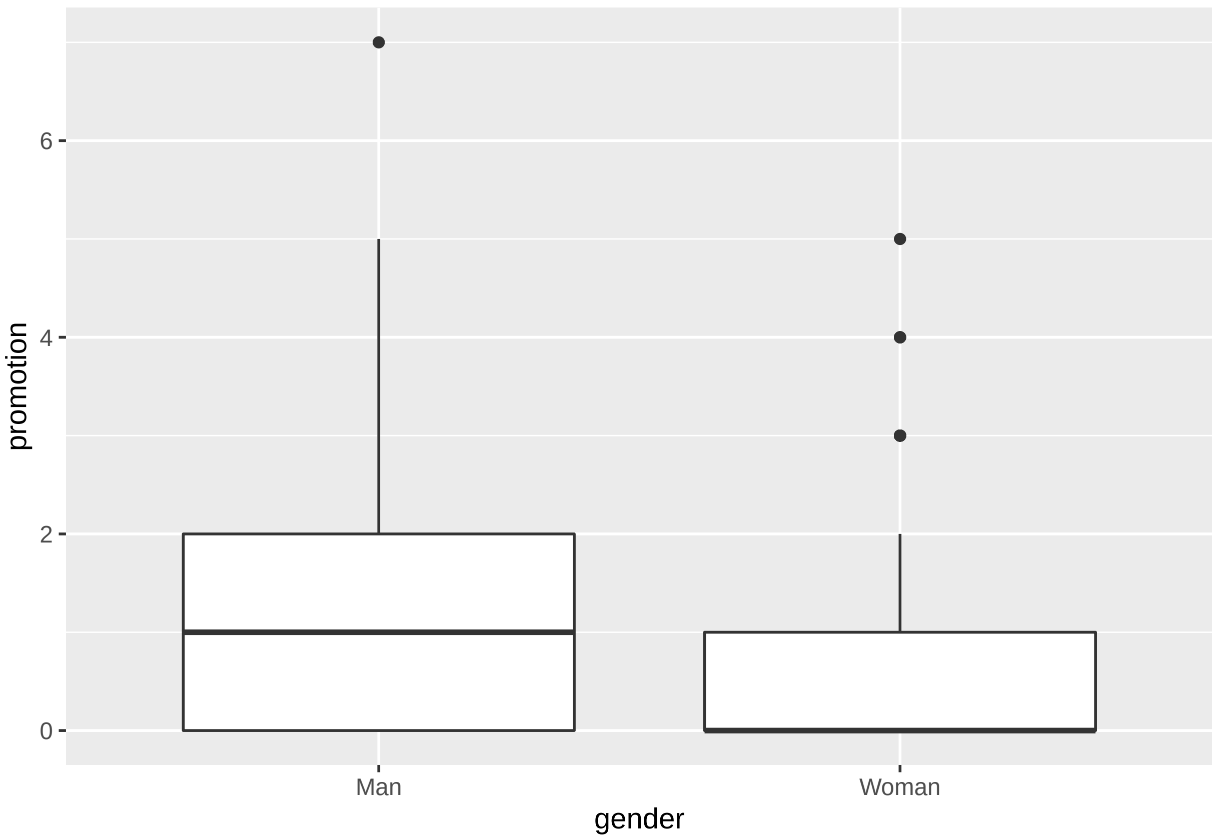


Figure 1: The Number of Promotion By Gender

In order to examine the potential bias in promotion, we derived a promotion data set from the data on current employees. We counted the number of times promoted for each empoyee and made inference on this count using gender. We filtered out those who prefer not to disclose their gender in order to improve the accuracy of inference, which is fine because there are 11 of them out of 608 employees. Figure 1 indicates the number of times being promoted across male and female employees. Observe that the meadian of promotion times for male is higher than the median for females, 1 for males and 0 for females. The employee being promoted the most is male and he got promoted 7 times.

We fitted a general linear model on the promotion data set using gender as a predictor for the number of times being promoted. The result indicates that female employees is promoted less than male employees, however(. . .)

Data Visualization

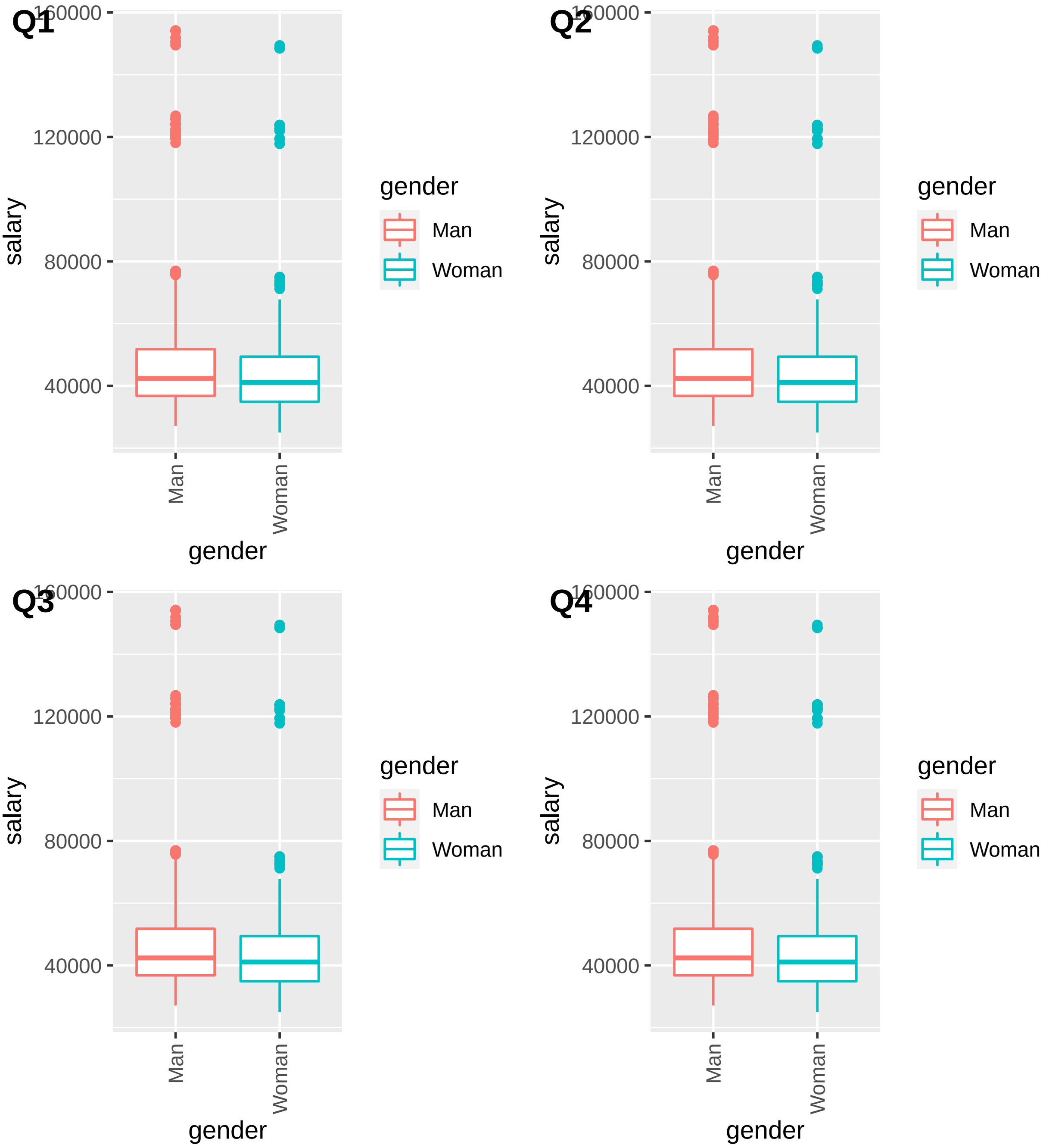


Figure 2: Salary Distribution for Men and Women in Each Quarter

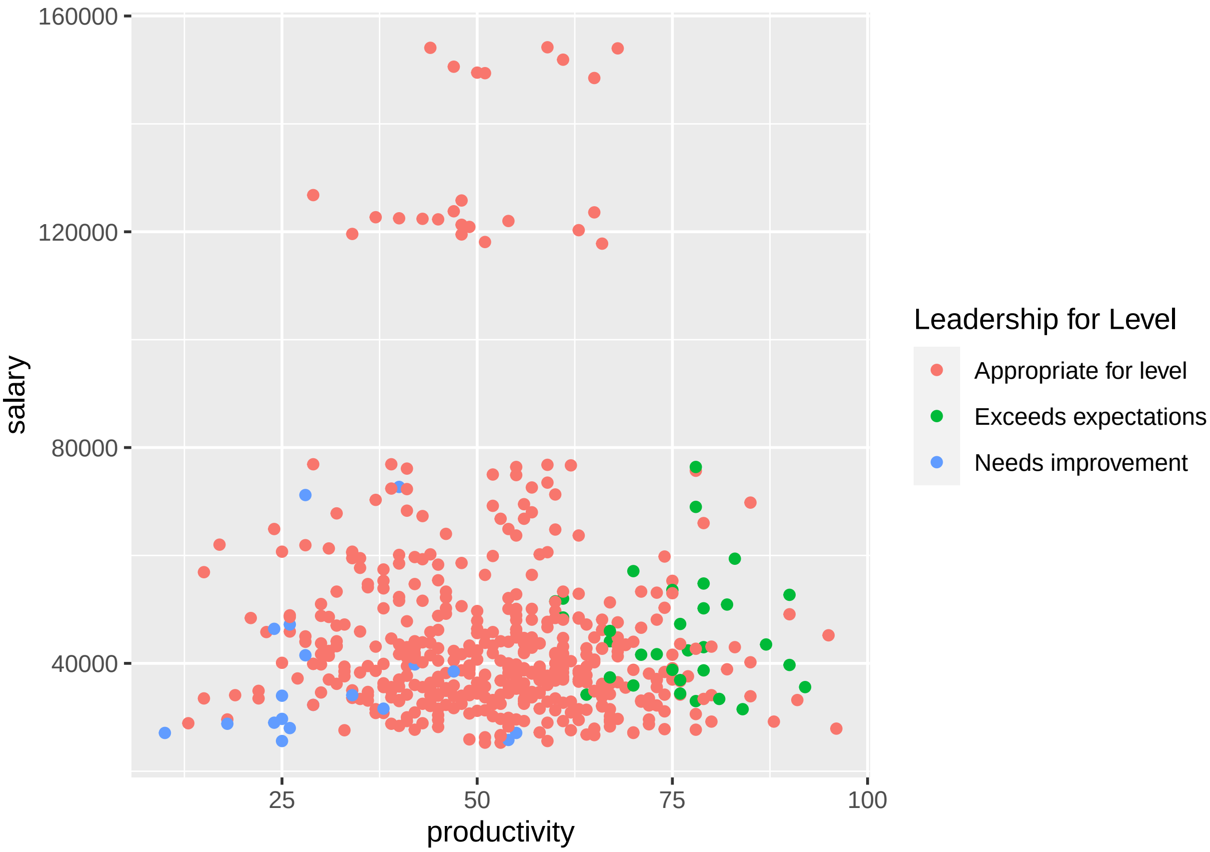


Figure 3: Salary Difference Across Leadership for Level Fixing Productivity

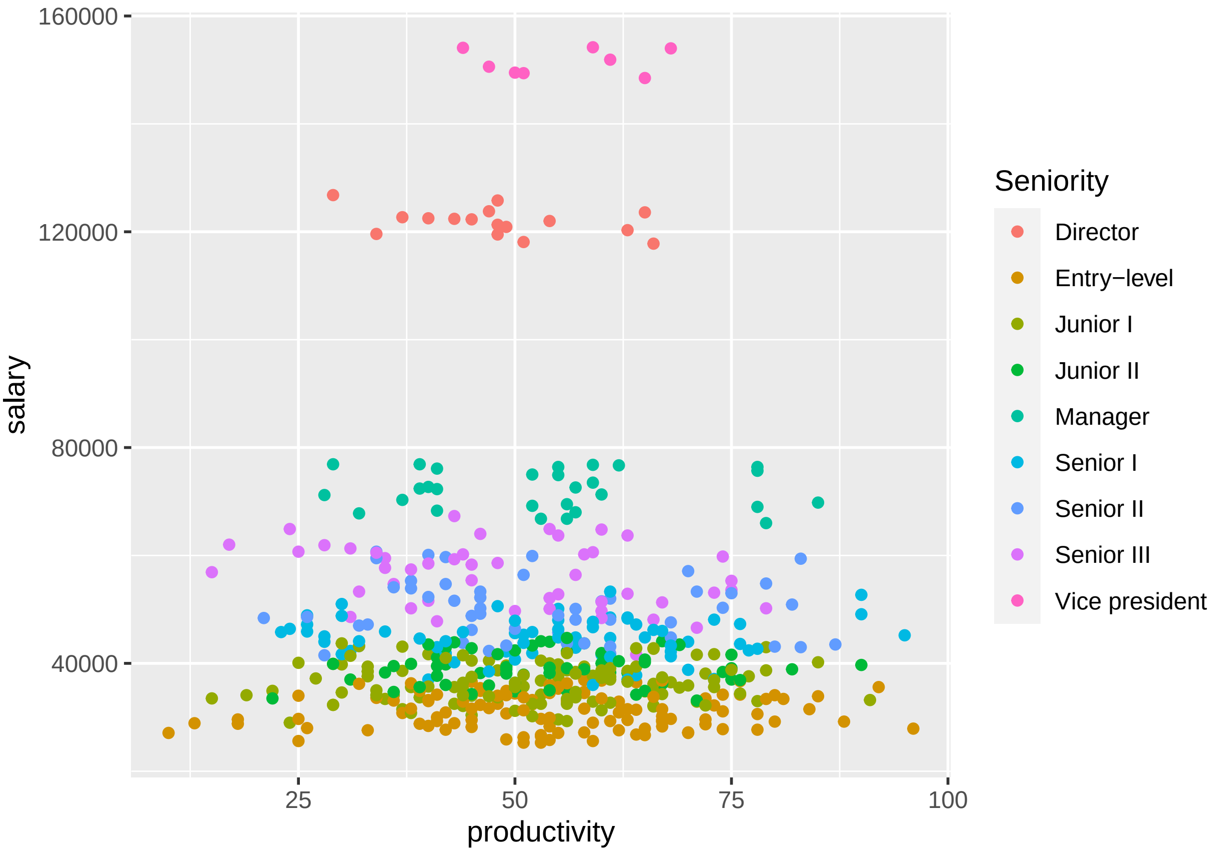


Figure 4: Salary Difference Across Role Seniority

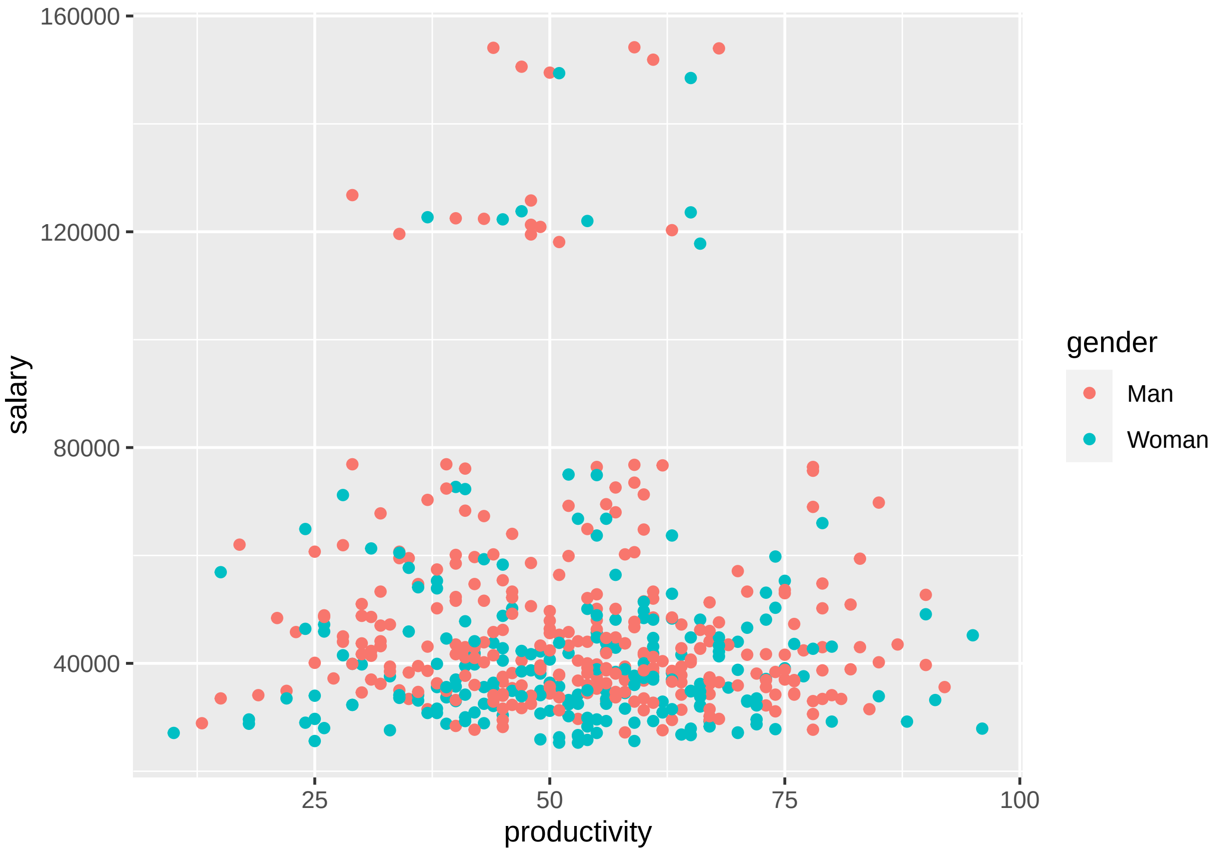


Figure 5: Salary Difference Across Leadership Levels

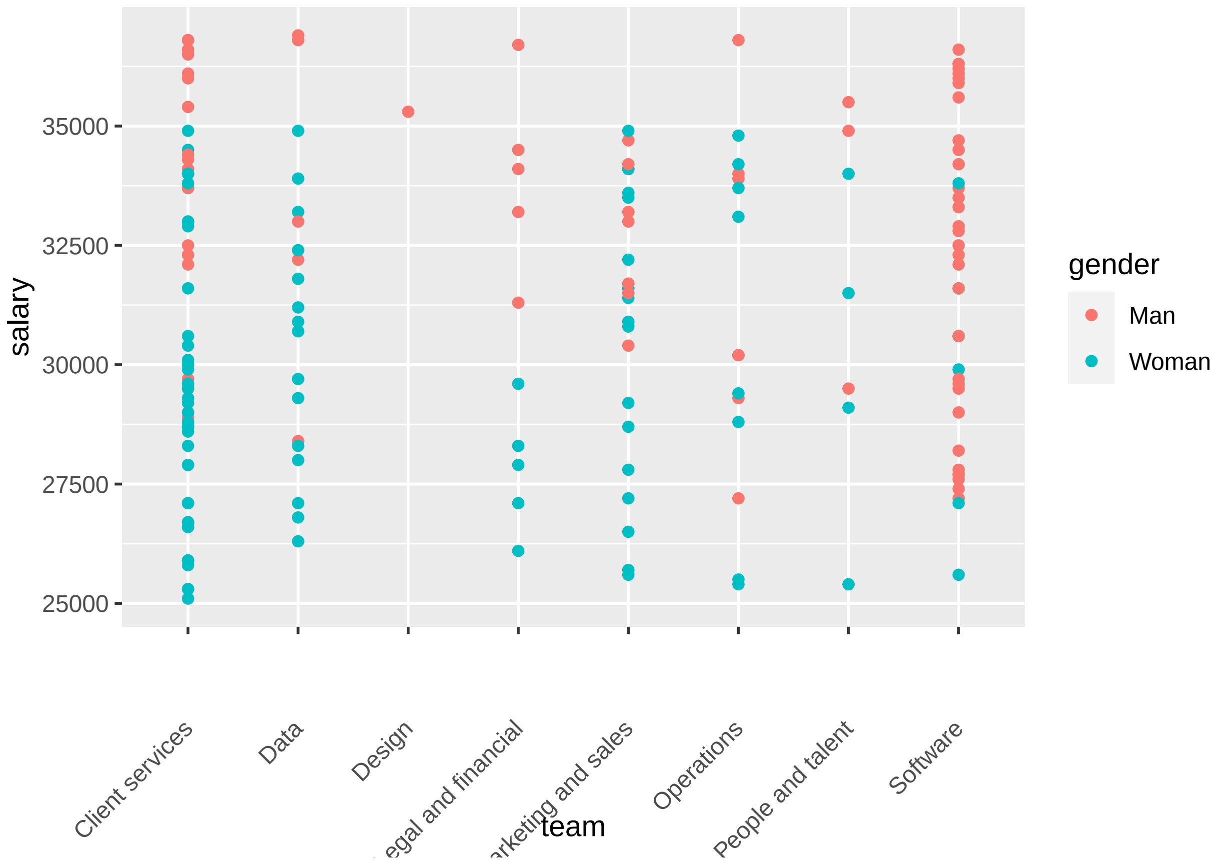


Figure 6: Salary Difference in Gender Across Teams, Fixing Quater and Seniority(Entry Level)

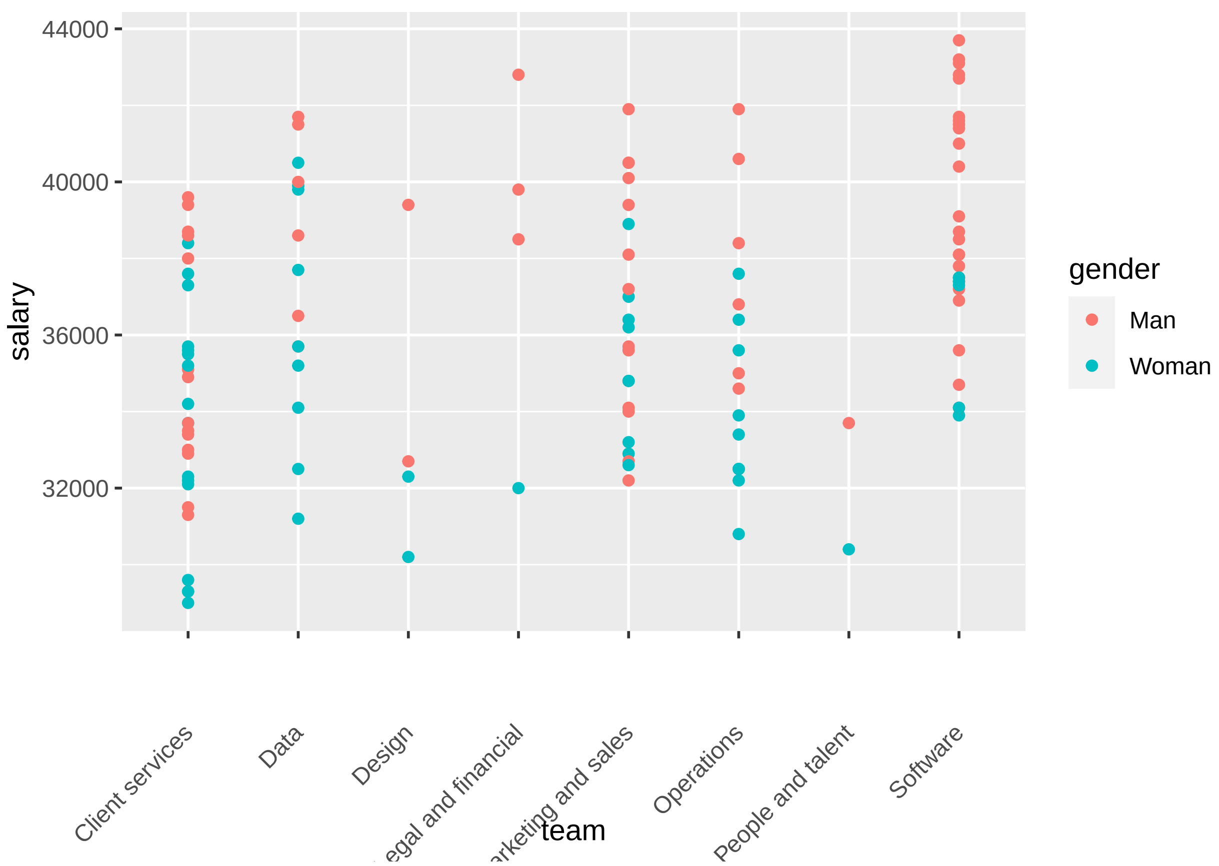


Figure 7: Salary Difference in Gender Across Teams, Fixing Quater and Seniority(Junior I)

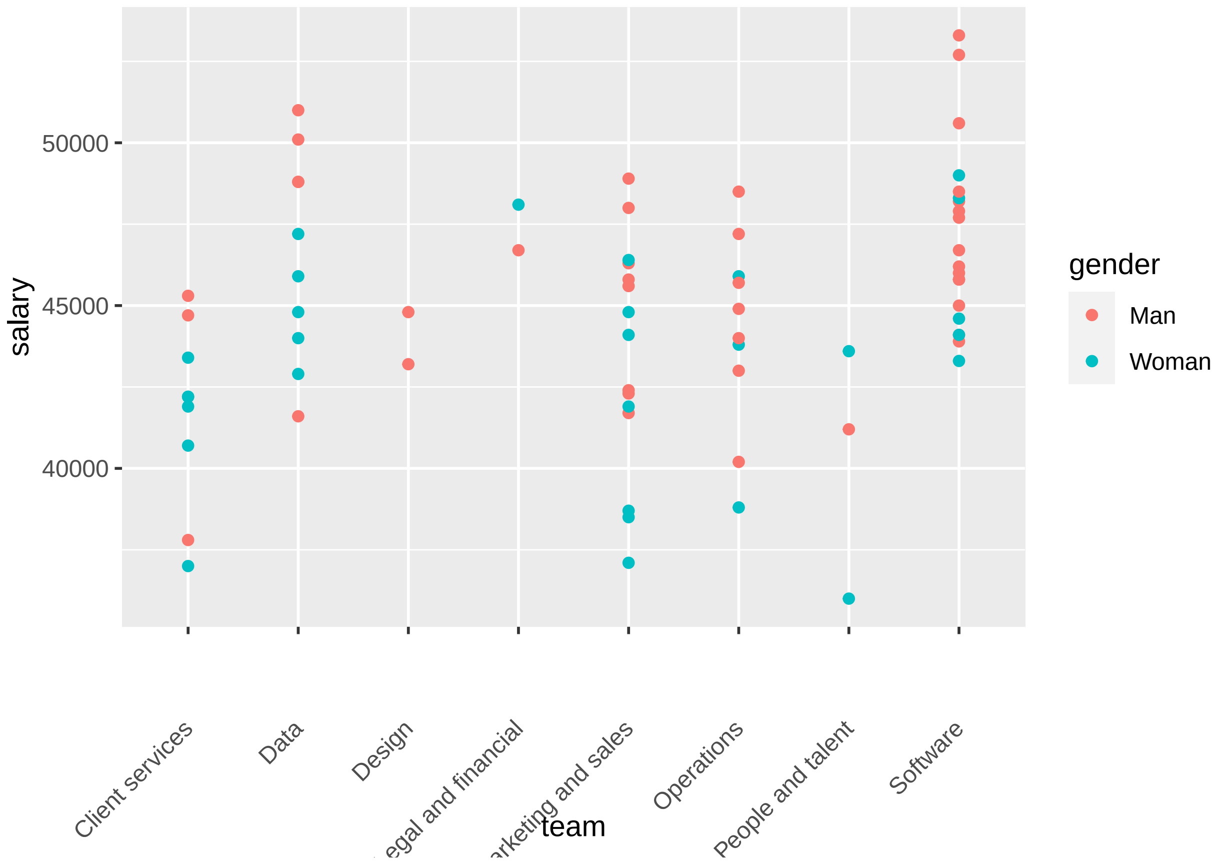


Figure 8: Salary Difference in Gender Across Teams, Fixing Quater and Seniority(Junior II)

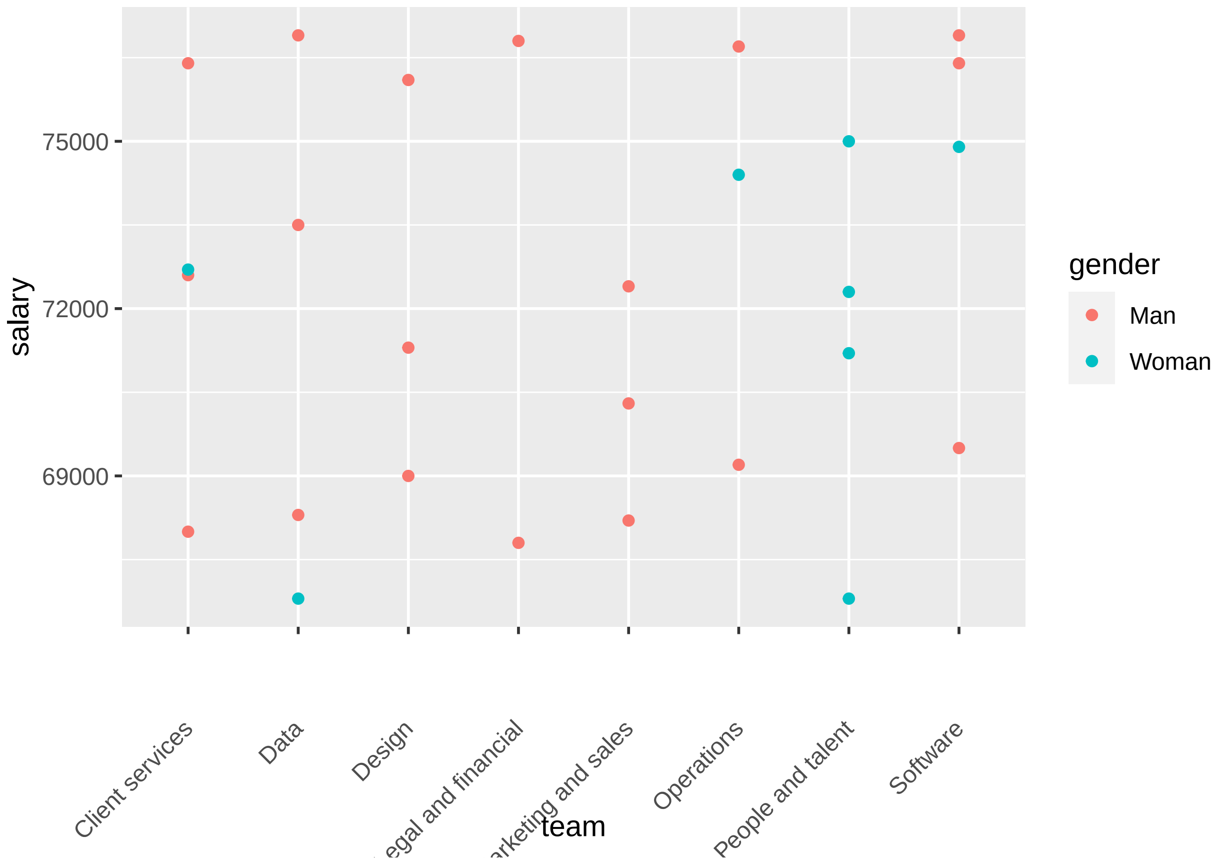


Figure 9: Salary Difference in Gender Across Teams, Fixing Quater and Seniority(Manager)

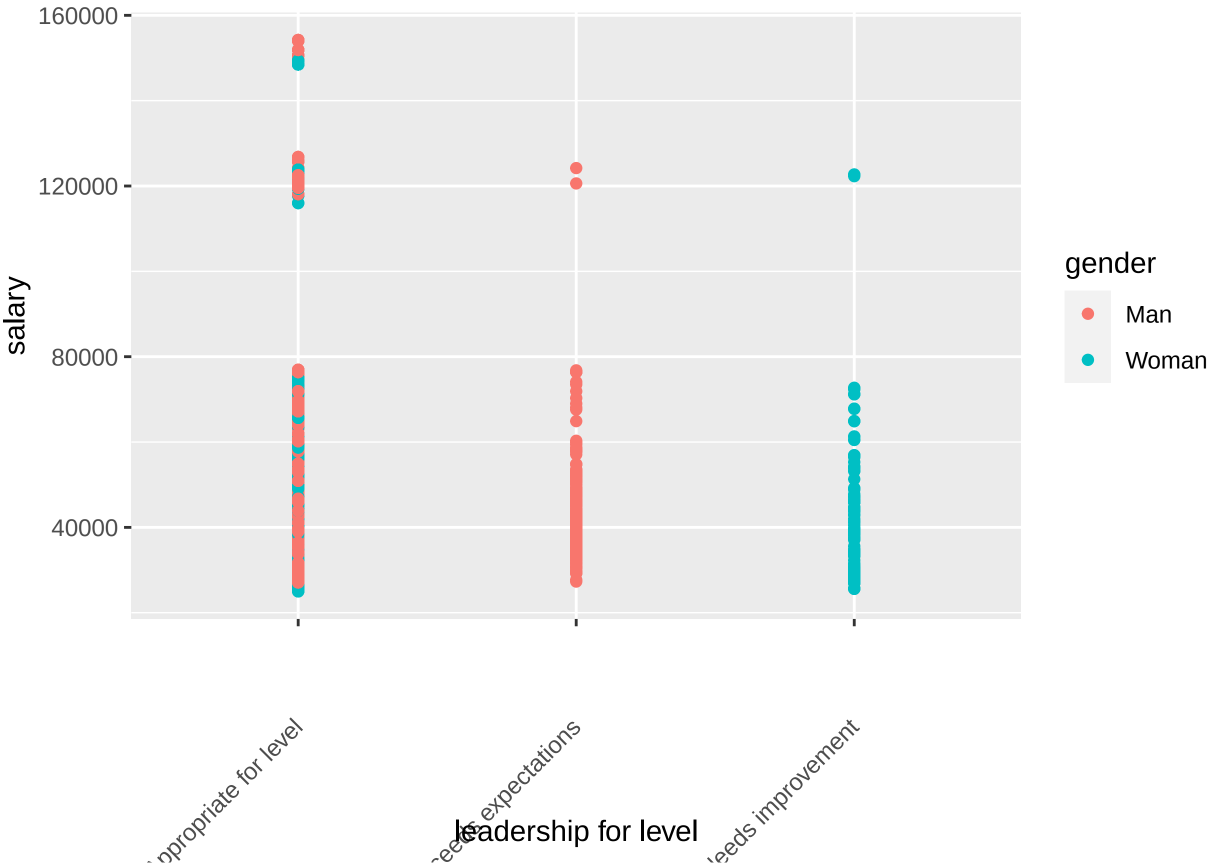


Figure 10: Salary Difference in Gender Across Leadership for Level

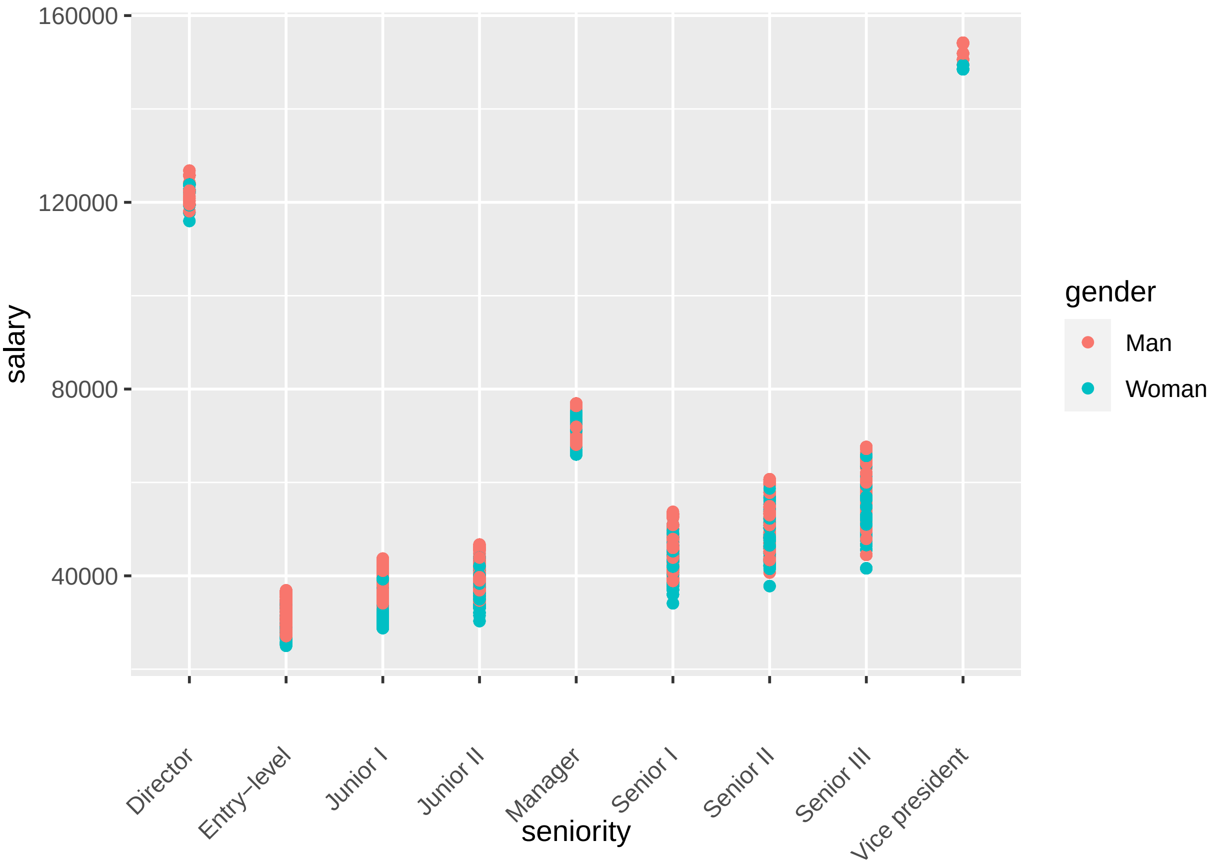


Figure 11: Salary Difference in Gender Across Seniority

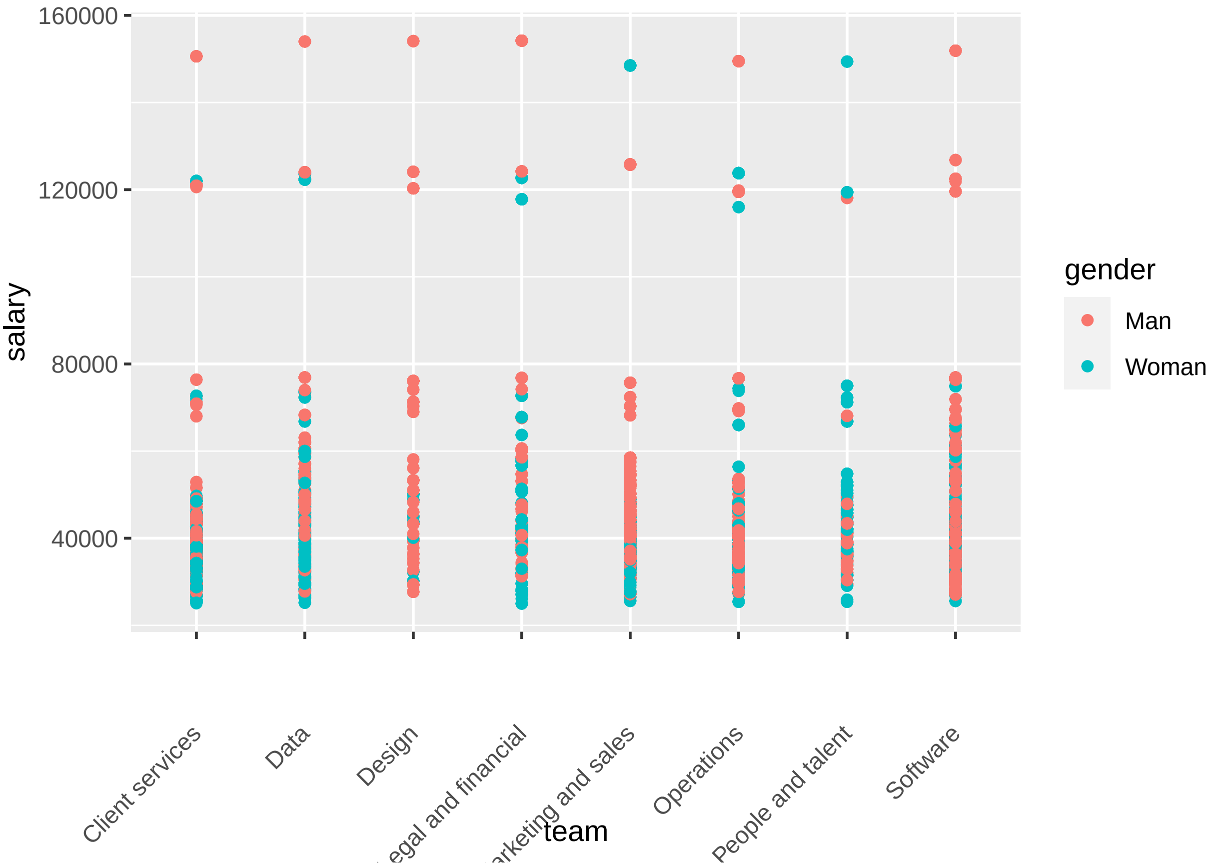


Figure 12: Salary Difference in Gender Across Teams

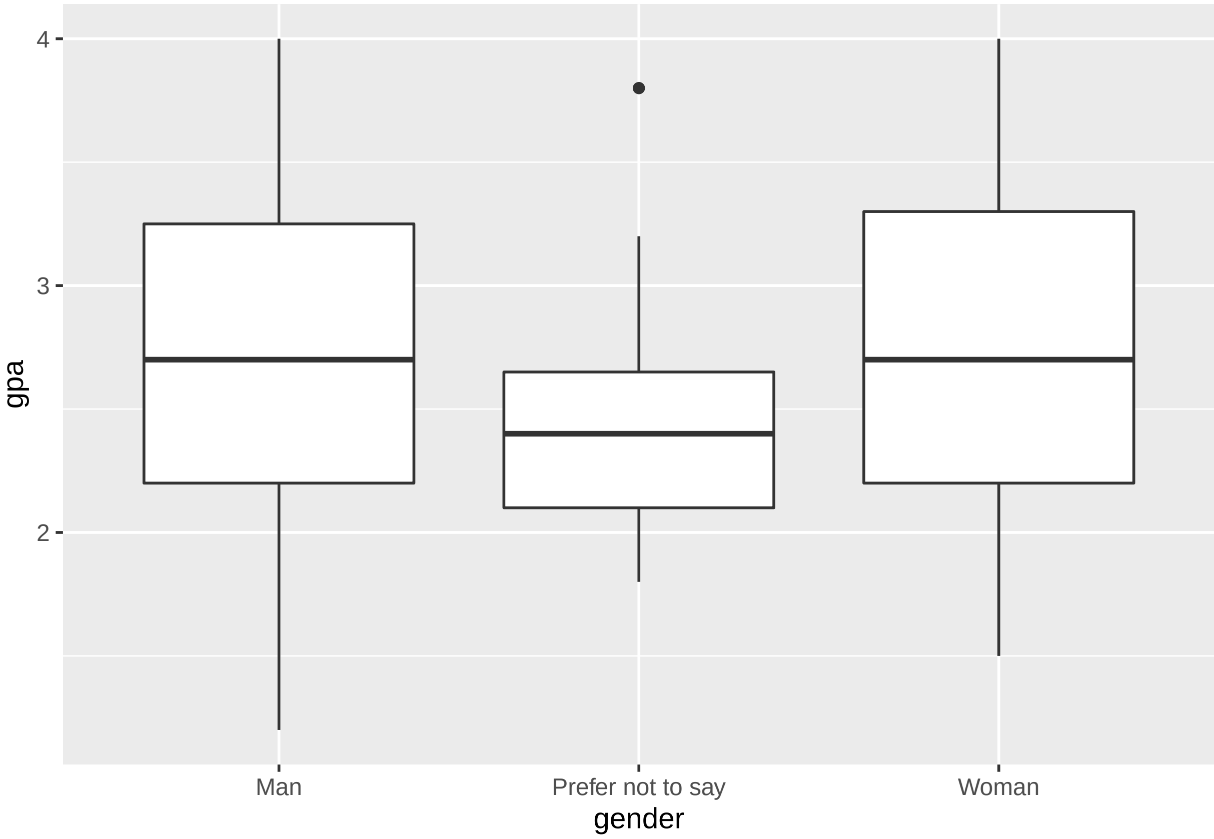
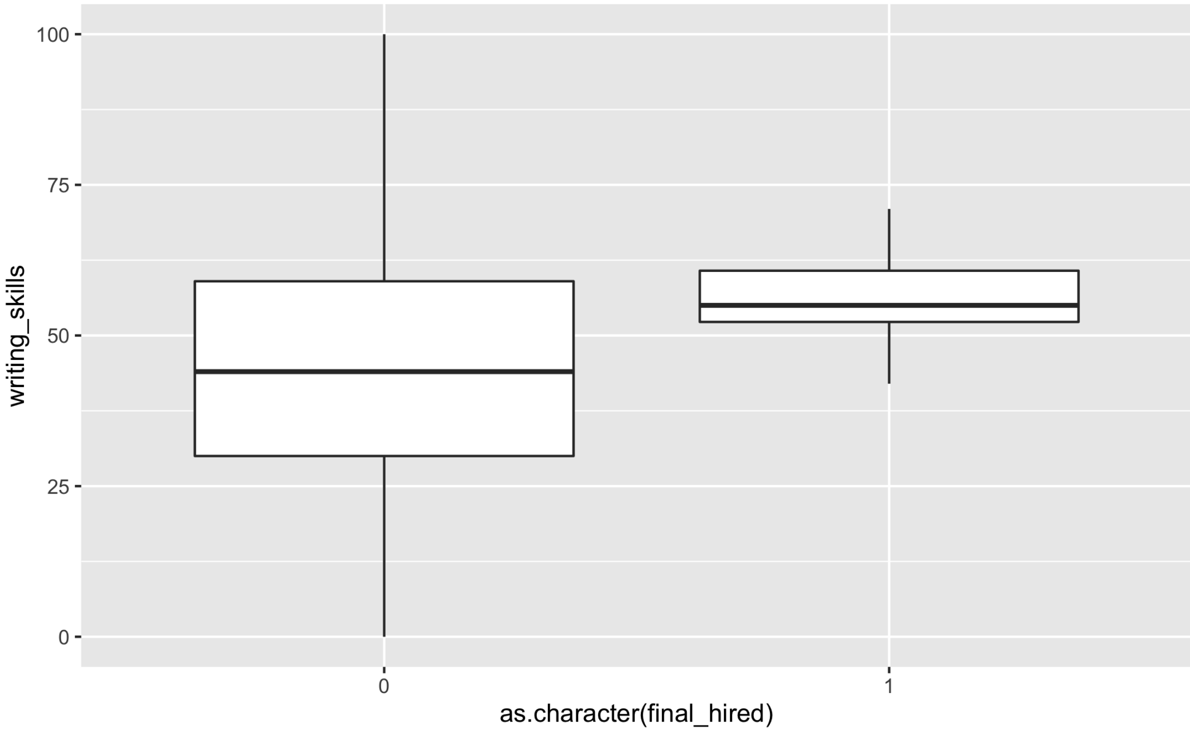


Figure 13: GPA Distribution Across Phase1 Applicants

```
## $x
## [1] "hiring results"
##
## attr(,"class")
## [1] "labels"
```

Writing Skills VS if hired



Discussion

In this section you will summarize your findings across all the research questions and discuss the strengths and limitations of your work. It doesn't have to be long, but keep in mind that often people will just skim the intro and the discussion of a document like this, so make sure it is useful as a semi-standalone section (doesn't have to be completely standalone like the executive summary).

Strengths and limitations

- The size of final hires data set is too small for inference. For data of this size(22 observations), inference of anykind is not meaningful. Alternatively, we rely on data visualization to come up with our conclusion.

Consultant information

Consultant profiles

Rain Wu. Rain is a senior consultant with DataOverFlow. She specializes in data visualization. Rain earned her Bachelor of Science, Specialist in Statistics Methods and Practice, from the University of Toronto in 2022. Before joining DataOverFlow, Rain has 3 year of working experience as a data engineer at Aviva in Markham, Toronto.

Tina Wang. Tina is a junior consultant with DataOverFlow. She specializes in reproducible analysis. Tina earned her Bachelor of Science, Majoring in Computer Science and Statistics from the University of Toronto in 2022. Tina earned her master degree in financial insurance from the University of Toronto in 2024.

Yiqu Ding. Yiqu is a junior consultant with DataOverFlow. She specializes in statistical communication. Yiqu earned her Bachelor of Science, Majoring in Statistics and mathmatical application in finance and economics from the University of Toronto in 2022. Yiqu earned her master degree in financial insurance from the University of Toronto in 2024.

Code of ethical conduct

- We respect and protect confidential data obtained from, or relating to, clients and third parties, as well as personal data and information about about employees from Data Over Flow. We only share information when there is a business purpose, and then do so in accordance with applicable laws and professional standards.
- We take proactive measures to safeguard our archives, computers and other data-storage devices containing confidential information or personal data. We promptly report any loss, damage or inapproprita disclosure of confidential information or personal data.
- We use social media and technology in a responsible way and respect everyone we work with. We obtain, develop and protect intellectual capital in an appropriate manner. We respect the resrtictions on its use and reproduction.