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 **3**

Technical report **4**

Introduction 4
Does there exist bias in current employee enumeration? 4
Does there exist bias in the hiring process? 4
Informative title for section addressing a research question 4
Data Visualization 5
Discussion 36

Consultant information **38**

Consultant profiles 38
Code of ethical conduct 38

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

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.

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 promition 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

Research questions

-
-
-

Does there exist bias in current employee enumeration?

For each research question, you will want to briefly describe any data manipulation, show some exploratory plots/summary tables, report on any methods you use (i.e. models you fit) and the conclusions you draw from these

Does there exist bias in the hiring process?

For each research question, you will want to briefly describe any data manipulation, show some exploratory plots/summary tables, report on any methods you use (i.e. models you fit) and the conclusions you draw from these

Informative title for section addressing a research question

For each research question, you will want to briefly describe any data manipulation, show some exploratory plots/summary tables, report on any methods you use (i.e. models you fit) and the conclusions you draw from these

Data Visualization

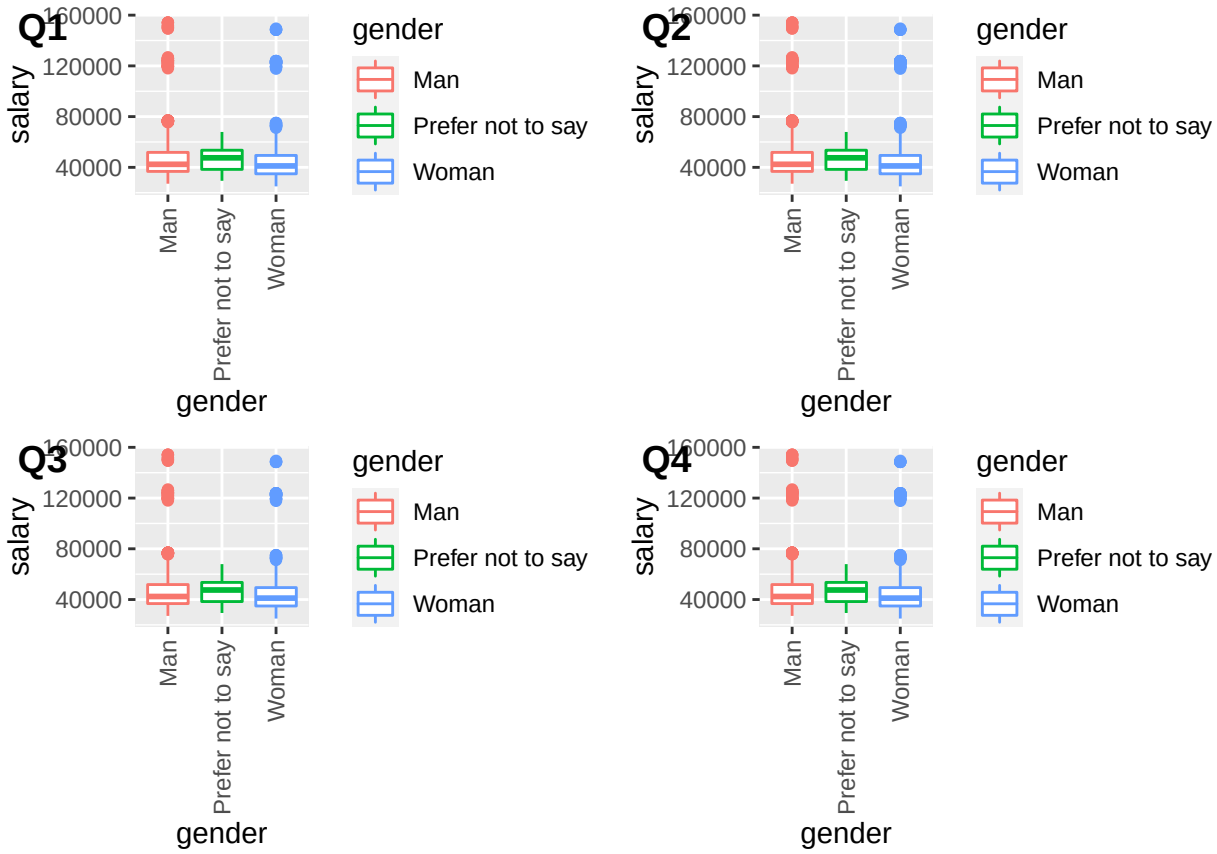
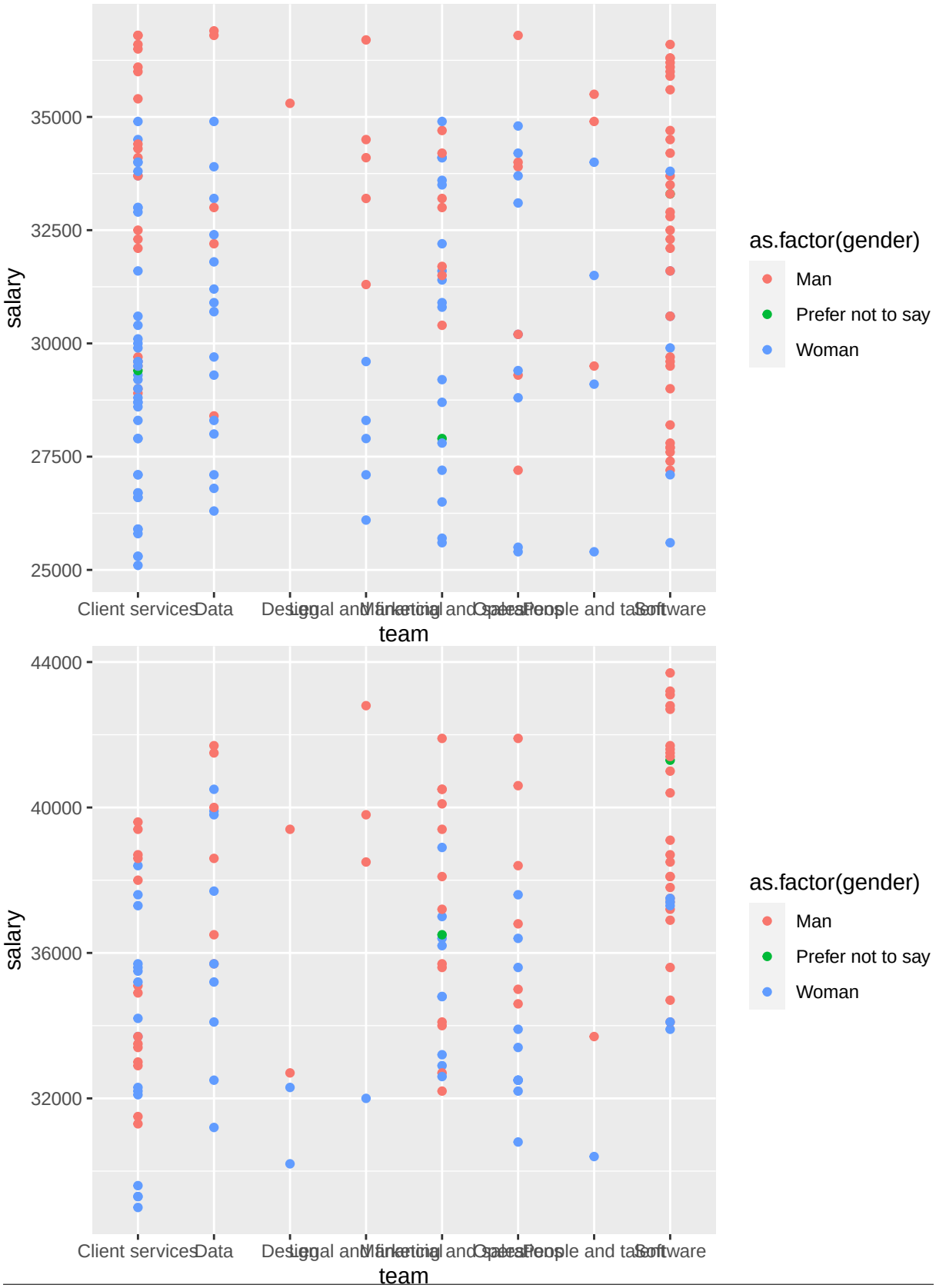
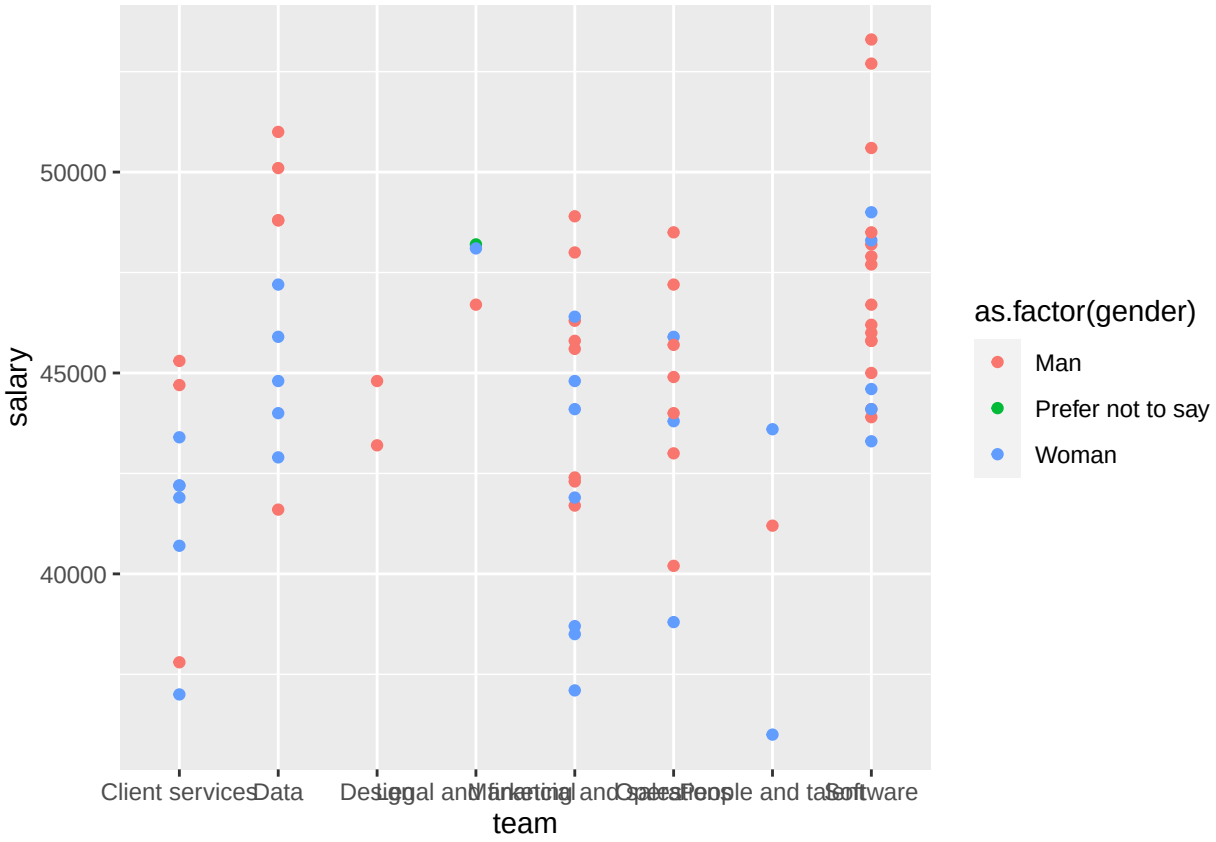
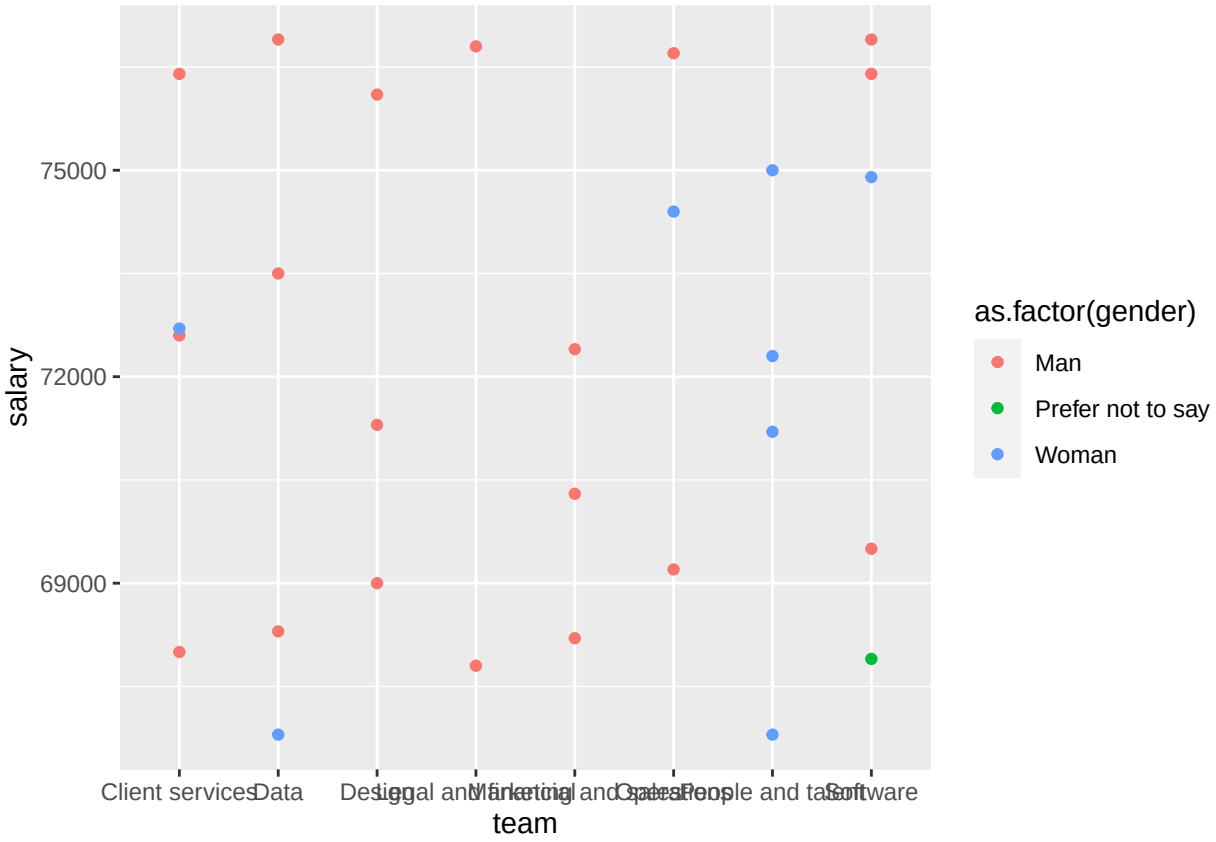


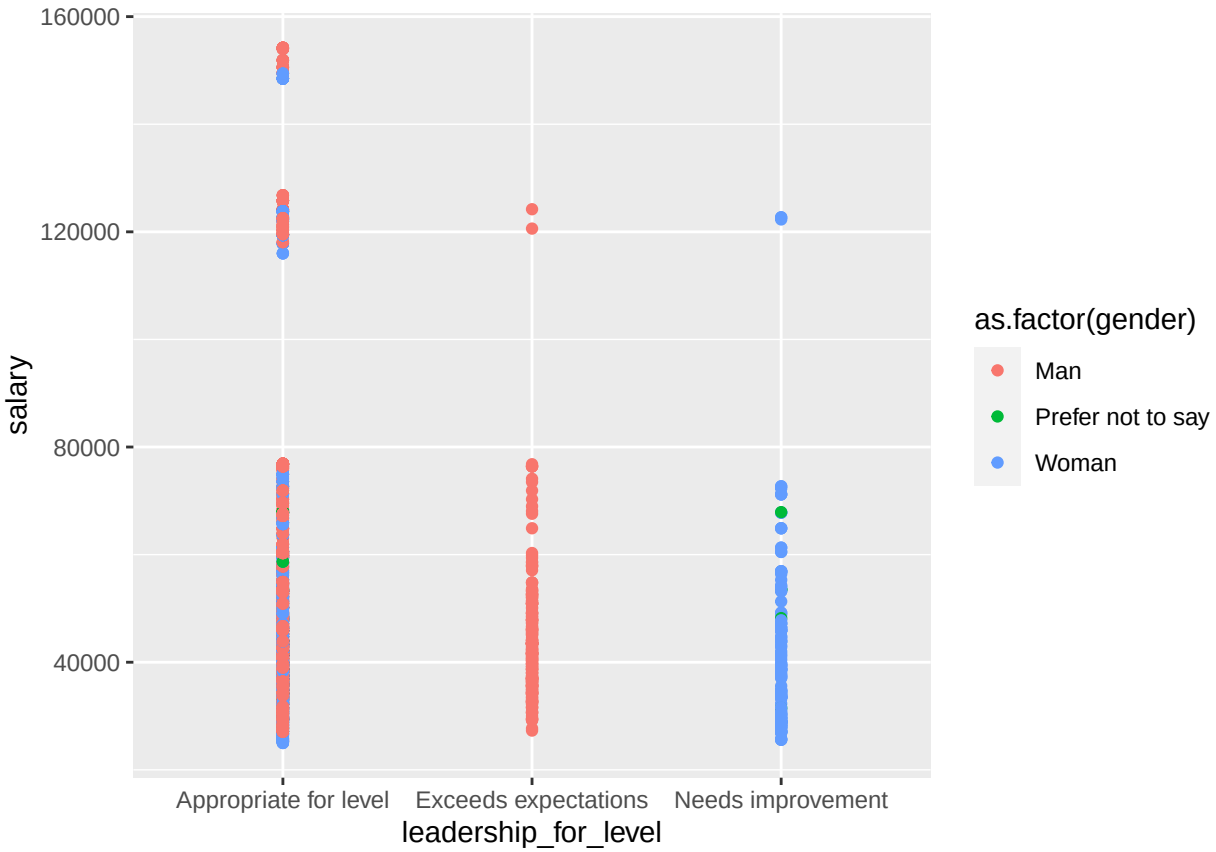
Figure 1: Salary Distribution for Men and Women in Each Quarter

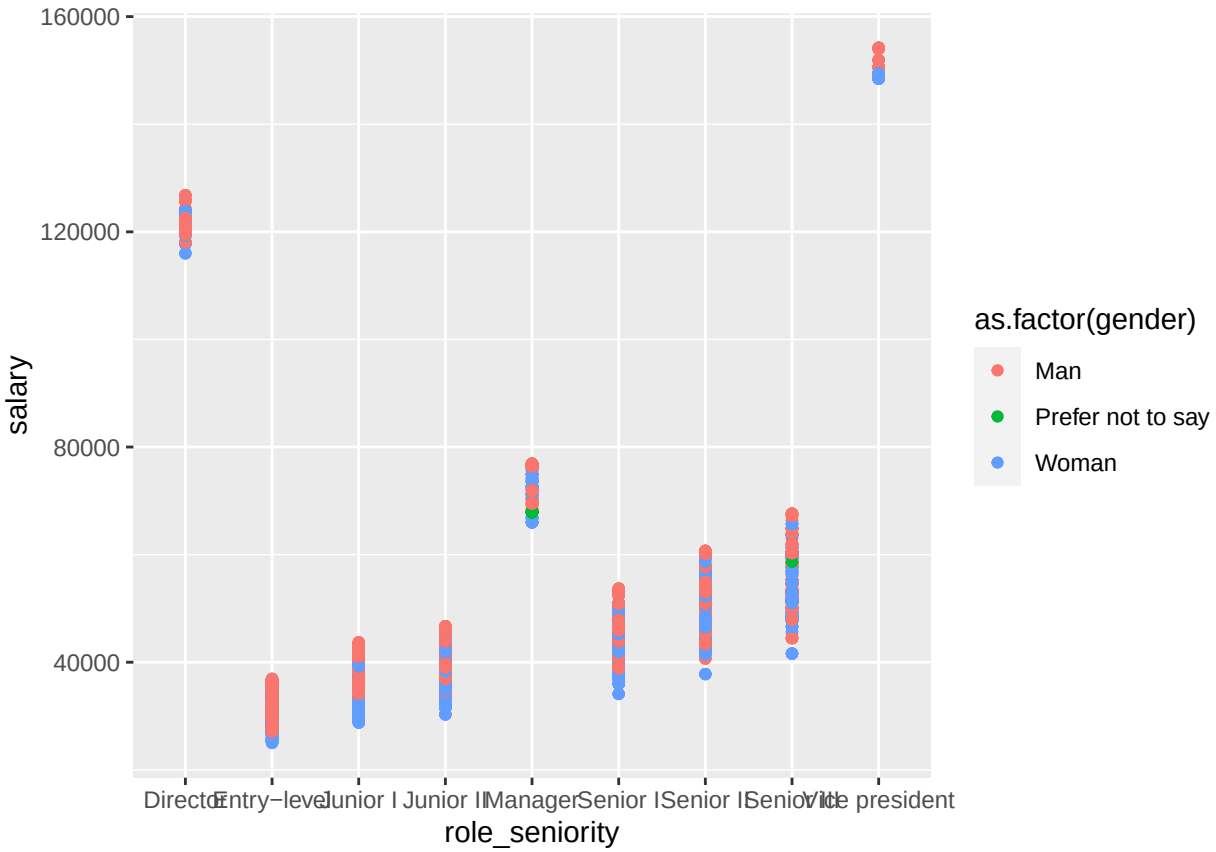
salary difference in gender across teams, fixing quarter and seniority

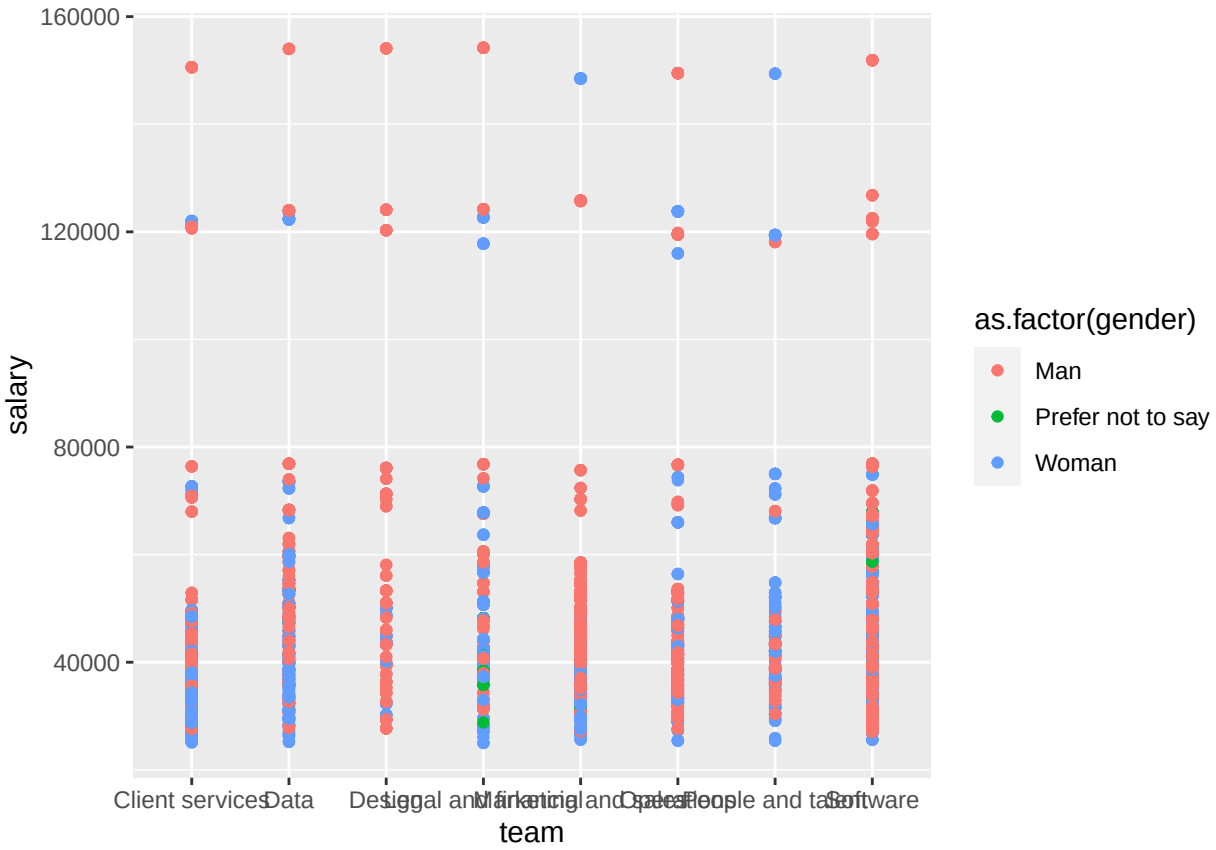



50000
45000
40000
salary
Client services
Data
Legal and
Marketing and
Sales
People and talent
Software
team
as.factor(gender)
Man
Prefer not to say
Woman


75000
72000
69000
salary
Client services
Data
Design
Legal and
Marketing
and
Sales
People and talent
Software
team
as.factor(gender)
Man
Prefer not to say
Woman


160000
120000
80000
40000
salary
Appropriate for level
Exceeds expectations
Needs improvement
leadership_for_level
as.factor(gender)
Man
Prefer not to say
Woman


160000
120000
80000
40000
salary
Director
Entry−level
Junior I
Junior II
Manager
Senior I
Senior II
Senior III
Vice president
role_seniority
as.factor(gender)
Man
Prefer not to say
Woman




```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salary ~ gender + role_seniority + financial_q + (1 | team) +
##     (1 | leadership_for_level) + (1 | productivity)
##    Data: current
##
## REML criterion at convergence: 131099.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2023 -0.7420  0.0259  0.7217  2.8209
##
## Random effects:
##  Groups               Name        Variance Std.Dev.
##  productivity         (Intercept)   175447  418.9
##  team                 (Intercept)  4728844 2174.6
##  leadership_for_level (Intercept)        0    0.0
##  Residual                         11012729 3318.5
```

```
## Number of obs: 6906, groups:
## productivity, 99; team, 8; leadership_for_level, 3
##
## Fixed effects:
##                               Estimate Std. Error  t value
## (Intercept)                  119786.22    3429.18   34.931
## genderPrefer not to say       -1370.63     316.28   -4.334
## genderWoman                   -1762.75      85.71  -20.566
## role_seniorityEntry-level    -91052.82     241.44 -377.120
## role_seniorityJunior I       -85669.12     236.80 -361.779
## role_seniorityJunior II      -83173.38     237.55 -350.128
## role_seniorityManager        -50779.17     277.90 -182.728
## role_seniority Senior I      -77882.04     241.05 -323.101
## role_seniority Senior II     -72455.57     243.13 -298.016
## role_seniority Senior III    -66868.66     248.75 -268.819
## role_seniorityVice president  28600.83     431.48   66.285
## financial_q2013 Q3              127.56    4706.39    0.027
## financial_q2013 Q4             1437.04    3721.55    0.386
## financial_q2014 Q1             2341.89    3471.22    0.675
## financial_q2014 Q2             2620.15    3450.62    0.759
## financial_q2014 Q3             2604.31    3409.62    0.764
## financial_q2014 Q4             2575.17    3395.82    0.758
## financial_q2015 Q1             3169.85    3374.06    0.939
## financial_q2015 Q2             3258.91    3362.35    0.969
## financial_q2015 Q3             3319.39    3353.31    0.990
## financial_q2015 Q4             2985.28    3350.28    0.891
## financial_q2016 Q1             3046.45    3348.70    0.910
## financial_q2016 Q2             2917.87    3346.72    0.872
## financial_q2016 Q3             2916.36    3345.68    0.872
## financial_q2016 Q4             2817.11    3344.86    0.842
## financial_q2017 Q1             2602.41    3343.91    0.778
## financial_q2017 Q2             2704.95    3342.78    0.809
## financial_q2017 Q3             2760.15    3341.78    0.826
## financial_q2017 Q4             2795.63    3341.90    0.837
## financial_q2018 Q1             2829.87    3340.91    0.847
## financial_q2018 Q2             2859.78    3340.39    0.856
## financial_q2018 Q3             2888.66    3339.90    0.865
## financial_q2018 Q4             2802.67    3339.66    0.839
```

```
## financial_q2019 Q1             2813.87    3339.18    0.843
## financial_q2019 Q2             2784.68    3339.20    0.834
## financial_q2019 Q3             2924.55    3338.78    0.876
## financial_q2019 Q4             2958.60    3338.67    0.886
## financial_q2020 Q1             3018.52    3338.20    0.904
## financial_q2020 Q2             3033.13    3338.16    0.909
## financial_q2020 Q3             3060.49    3337.88    0.917
## financial_q2020 Q4             3029.47    3337.77    0.908
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular

## Linear mixed model fit by REML ['lmerMod']
## Formula: salary ~ gender + role_seniority + financial_q + (1 | team) +
##     (1 | leadership_for_level)
##    Data: current
##
## REML criterion at convergence: 131126.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1863 -0.7558  0.0129  0.7257  2.8045
##
## Random effects:
##  Groups               Name        Variance Std.Dev.
##  team                 (Intercept)  4730675 2175.01
##  leadership_for_level (Intercept)     6499   80.61
##  Residual                          11154062 3339.77
## Number of obs: 6906, groups:  team, 8; leadership_for_level, 3
##
## Fixed effects:
##                                Estimate Std. Error  t value
## (Intercept)                   1.200e+05  3.440e+03   34.886
## genderPrefer not to say      -1.289e+03  3.170e+02   -4.067
## genderWoman                  -1.786e+03  8.593e+01  -20.783
## role_seniorityEntry-level    -9.110e+04  2.403e+02 -379.117
## role_seniorityJunior I       -8.572e+04  2.355e+02 -363.925
## role_seniorityJunior II      -8.321e+04  2.367e+02 -351.495
## role_seniorityManager        -5.075e+04  2.756e+02 -184.170
```

```
## role_senioritySenior I       -7.793e+04  2.398e+02 -325.036
## role_senioritySenior II      -7.245e+04  2.427e+02 -298.581
## role_senioritySenior III     -6.690e+04  2.475e+02 -270.267
## role_seniorityVice president  2.851e+04  4.313e+02   66.106
## financial_q2013 Q3           -1.149e-08  4.723e+03    0.000
## financial_q2013 Q4            1.355e+03  3.738e+03    0.362
## financial_q2014 Q1            2.142e+03  3.482e+03    0.615
## financial_q2014 Q2            2.449e+03  3.463e+03    0.707
## financial_q2014 Q3            2.462e+03  3.420e+03    0.720
## financial_q2014 Q4            2.415e+03  3.407e+03    0.709
## financial_q2015 Q1            3.048e+03  3.385e+03    0.901
## financial_q2015 Q2            3.177e+03  3.373e+03    0.942
## financial_q2015 Q3            3.170e+03  3.364e+03    0.942
## financial_q2015 Q4            2.834e+03  3.361e+03    0.843
## financial_q2016 Q1            2.894e+03  3.360e+03    0.862
## financial_q2016 Q2            2.741e+03  3.358e+03    0.816
## financial_q2016 Q3            2.747e+03  3.356e+03    0.818
## financial_q2016 Q4            2.674e+03  3.356e+03    0.797
## financial_q2017 Q1            2.463e+03  3.355e+03    0.734
## financial_q2017 Q2            2.593e+03  3.354e+03    0.773
## financial_q2017 Q3            2.643e+03  3.353e+03    0.788
## financial_q2017 Q4            2.671e+03  3.353e+03    0.797
## financial_q2018 Q1            2.718e+03  3.352e+03    0.811
## financial_q2018 Q2            2.743e+03  3.351e+03    0.818
## financial_q2018 Q3            2.728e+03  3.351e+03    0.814
## financial_q2018 Q4            2.660e+03  3.350e+03    0.794
## financial_q2019 Q1            2.700e+03  3.350e+03    0.806
## financial_q2019 Q2            2.694e+03  3.350e+03    0.804
## financial_q2019 Q3            2.798e+03  3.349e+03    0.835
## financial_q2019 Q4            2.840e+03  3.349e+03    0.848
## financial_q2020 Q1            2.880e+03  3.349e+03    0.860
## financial_q2020 Q2            2.915e+03  3.349e+03    0.870
## financial_q2020 Q3            2.928e+03  3.349e+03    0.875
## financial_q2020 Q4            2.896e+03  3.348e+03    0.865

## Likelihood ratio test
##
## Model 1: salary ~ gender + role_seniority + financial_q + (1 | team) +
```

```
##     (1 | leadership_for_level) + (1 | productivity)
## Model 2: salary ~ gender + role_seniority + financial_q + (1 | team) +
##     (1 | leadership_for_level)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  45 -65550
## 2  44 -65563 -1 26.917  2.124e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: salary ~ gender + role_seniority + financial_q + (1 | team) +
##     (1 | leadership_for_level)
## Model 2: salary ~ gender + role_seniority + financial_q + (1 | leadership_for_level)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  44 -65563
## 2  43 -66528 -1 1930.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: salary ~ gender + role_seniority + financial_q + (1 | leadership_for_level)
## Model 2: salary ~ gender + role_seniority + financial_q + (1 | team)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  43 -66528
## 2  43 -65563  0 1930.1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part 2 on hiring data findings: 1.men and women have similar GPA 2.applicants with higher GPA are hired 3.applicants with better skills are hired 4.largest difference in speaking skills, and least in minimal skills -may suggest bias towards non-native speakers 5.on average, female hires have lower skills than male hires

#research question
-is there gender bias in the hiring process -model: #create phase 2 and phase 3 hired by merging -phase2_hired ~ gender* cv gpa cover_letter (phase1) -phase3_hired ~ gender* cv gpa cover_letter tech writing speaking (phase2) -final_hired ~ gender* cv gpa cover_letter tech

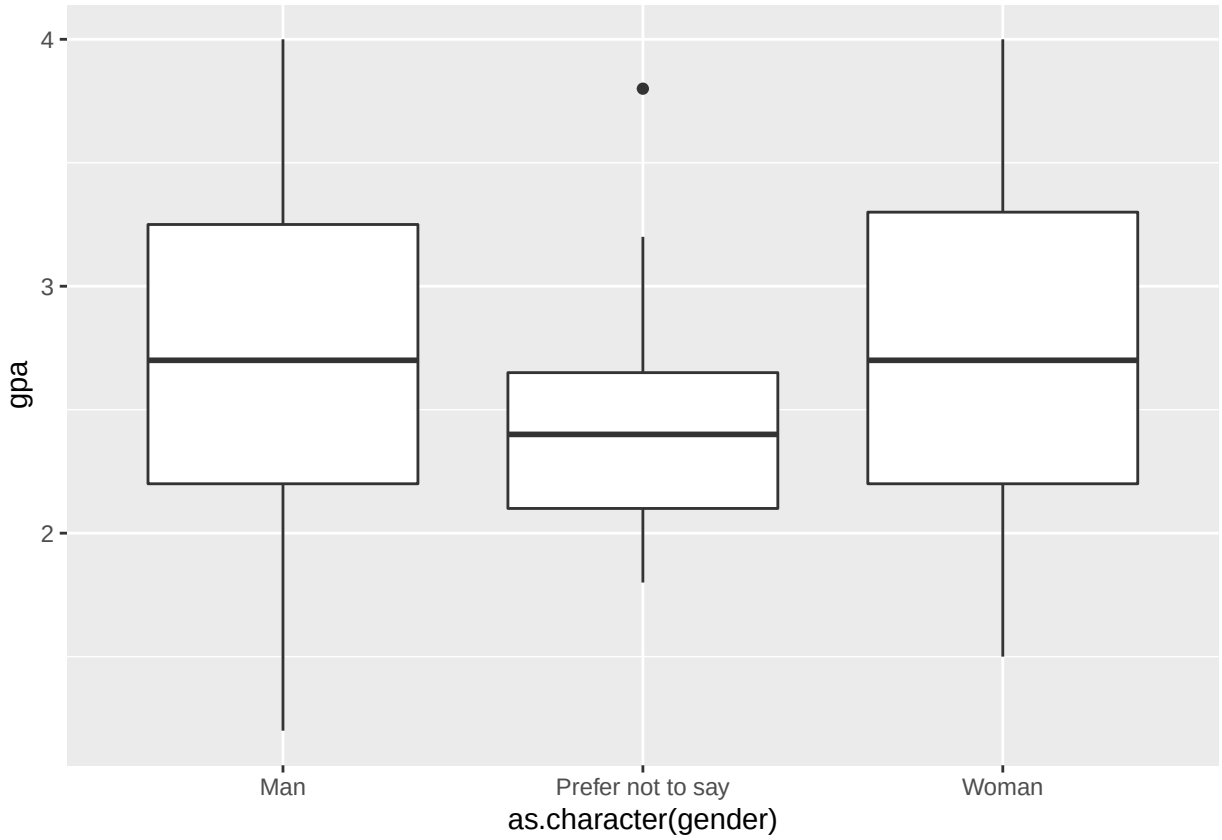
writing speaking rating1 rating2 (phase3) -y 0 or 1, not continous -not linear reg or linear mixed -need generalized linear model -all fixed effect

-reduced model
-final_hired ~ gender gpa tech writing speaking
-model comparison

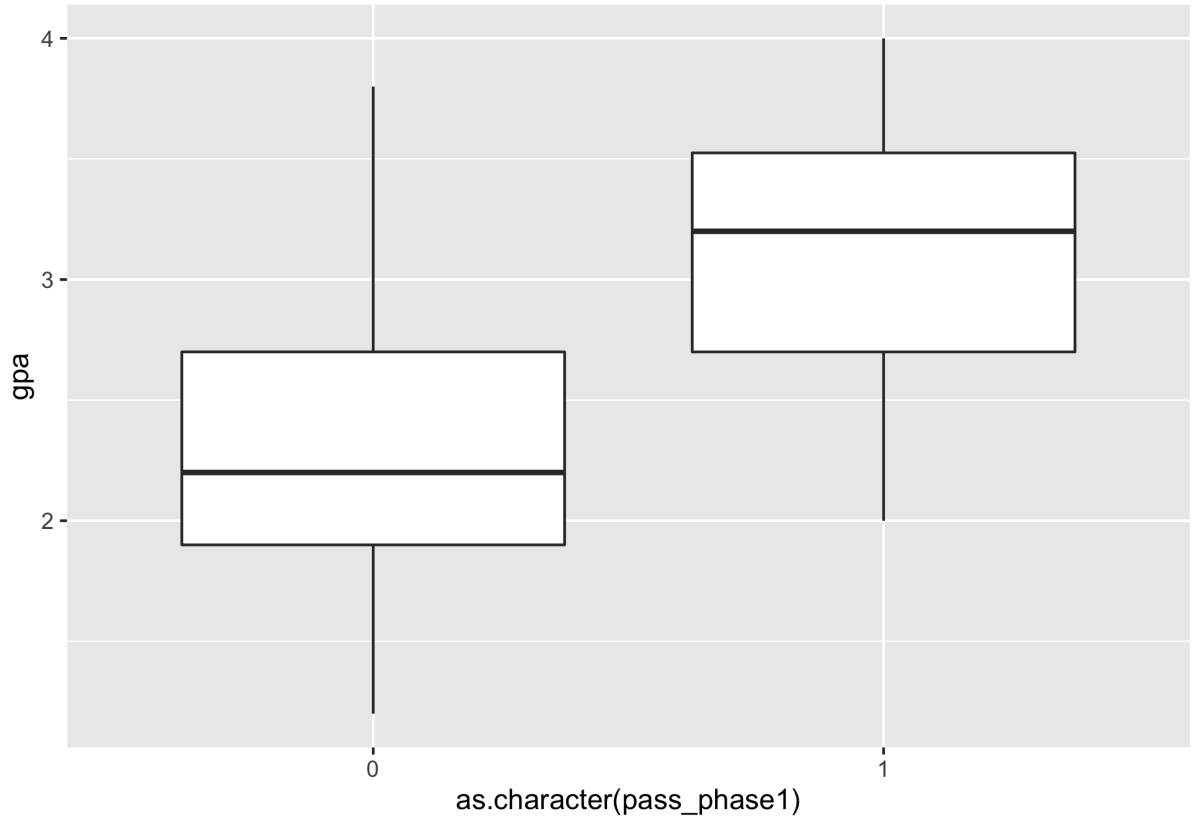
-model
-gpa ~ gender
-skills ~ gender

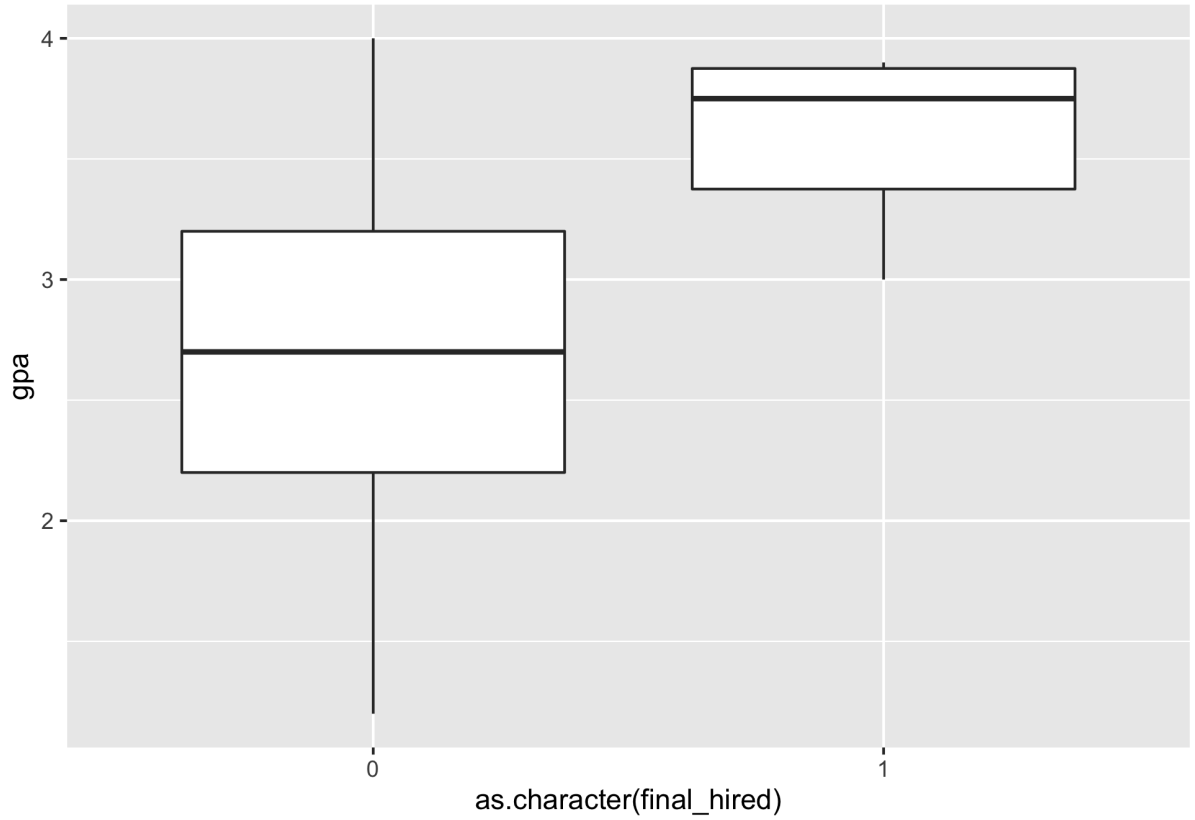
-is there race bias in the hiring process -phase3_hired ~ speaking* gender cv gpa cover_letter tech writing (phase2) -final_hired ~ speaking* gender cv gpa cover_letter tech writing rating1 rating2 (phase3)

```
## # A tibble: 300 x 15
##       X1 applicant_id team_applied_for cover_letter    cv   gpa gender
##    <dbl>        <dbl> <chr>                   <dbl> <dbl> <dbl> <chr>
##  1     1         1030 Data                        1     1   2.4 Woman
##  2     2         1070 Software                    1     1   3.4 Woman
##  3     3         1080 Data                        1     1   2.6 Man
##  4     4         1090 Data                        1     1   3.7 Man
##  5     5         1120 Software                    1     1   3.8 Woman
##  6     6         1140 Data                        1     1   3.3 Woman
##  7     7         1150 Software                    1     1   3.2 Man
##  8     8         1170 Software                    1     1   2.9 Man
##  9     9         1180 Software                    1     1   3.2 Man
## 10    10         1230 Software                    1     1   2.8 Man
## # ... with 290 more rows, and 8 more variables: extracurriculars <dbl>,
## #   work_experience <dbl>, technical_skills <dbl>, writing_skills <dbl>,
## #   leadership_presence <dbl>, speaking_skills <dbl>, final_hired <dbl>,
## #   pass_phase2 <dbl>
```

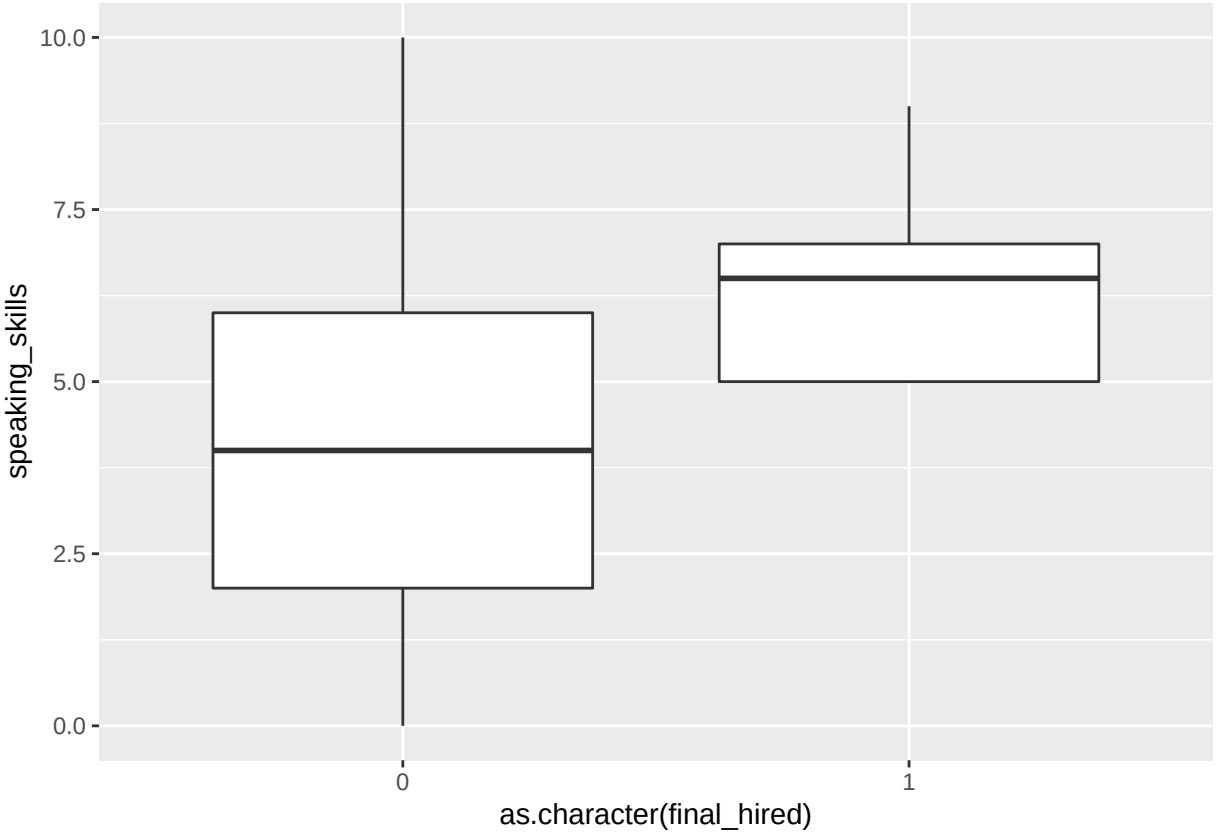

4
3
2
gpa
Man
Prefer not to say
Woman
as.character(gender)

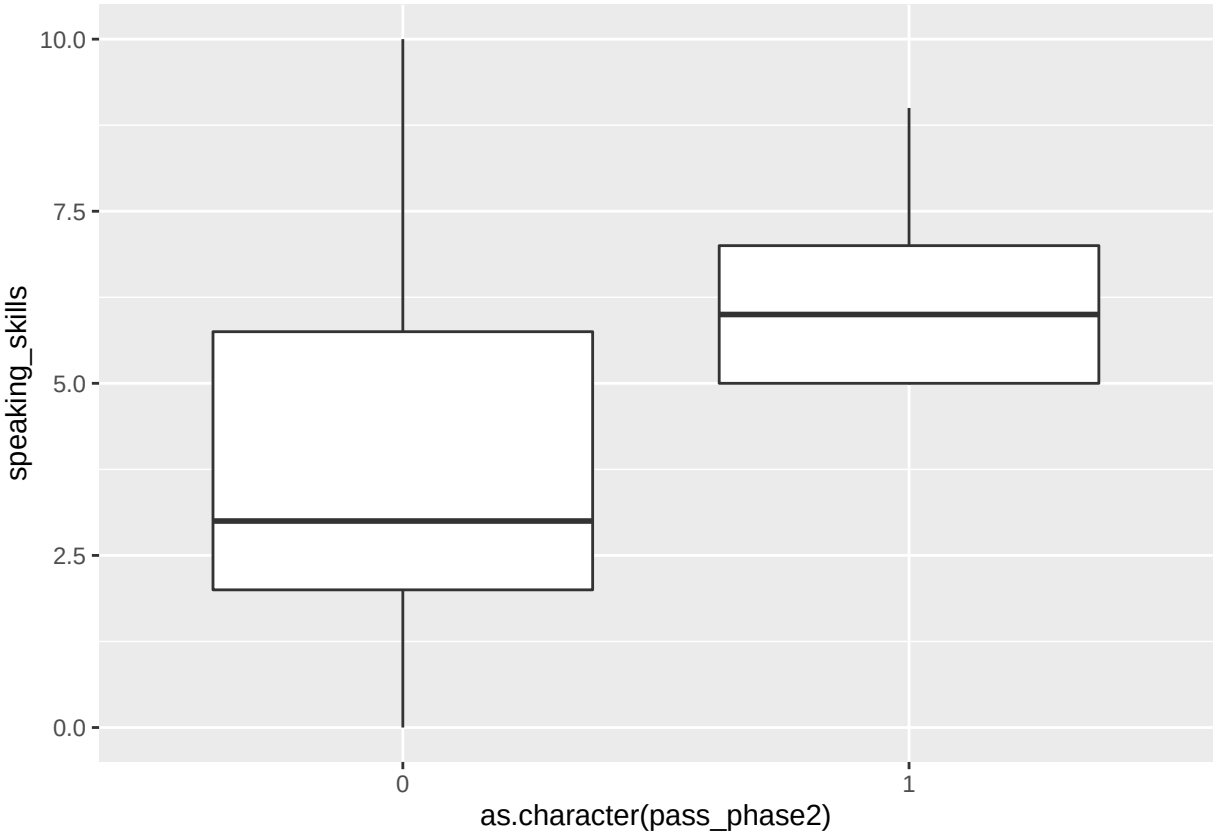
GPA vs. if hired

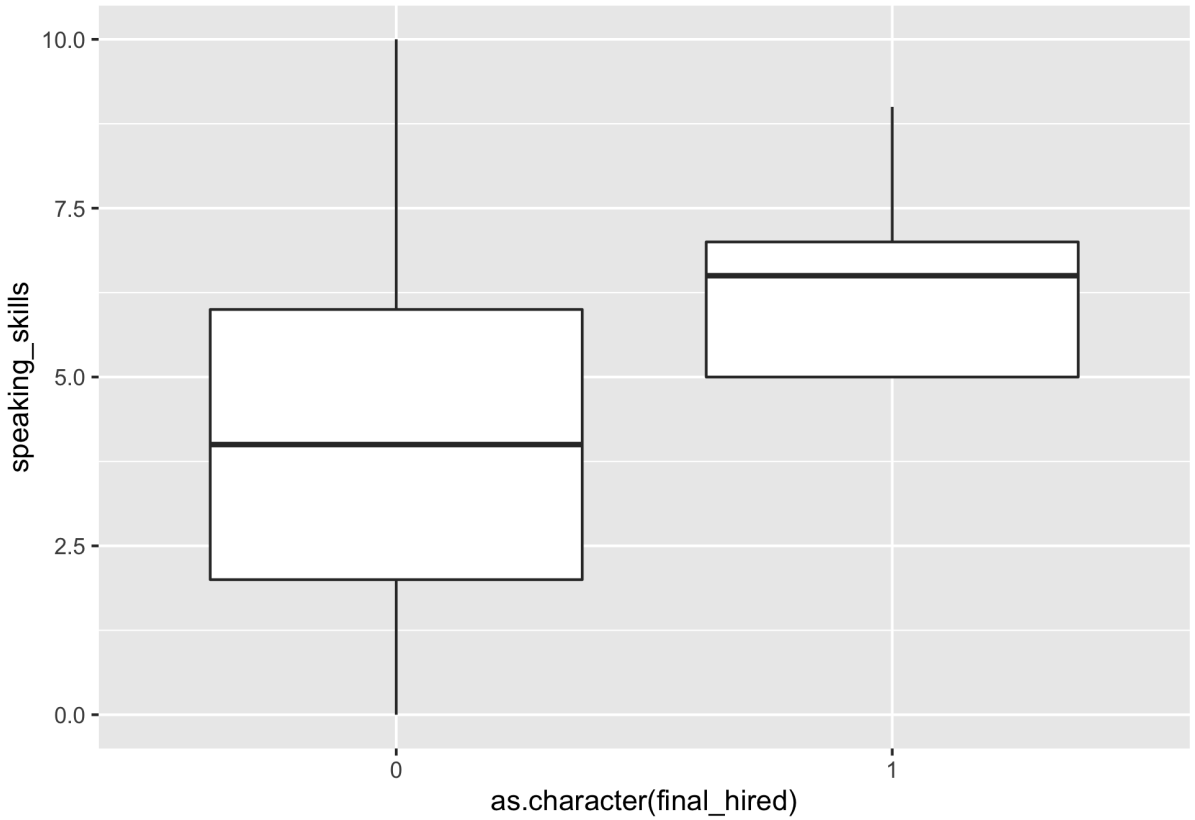


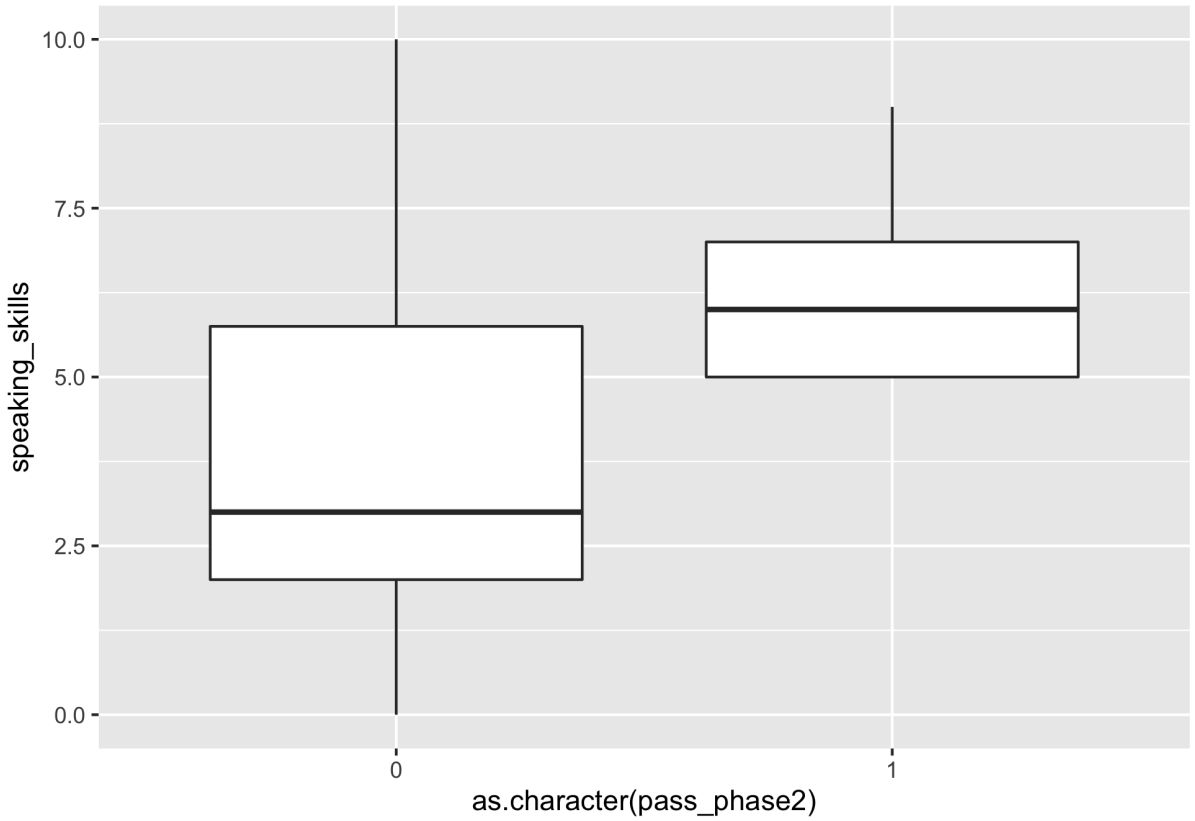


Speaking skills VS if hired

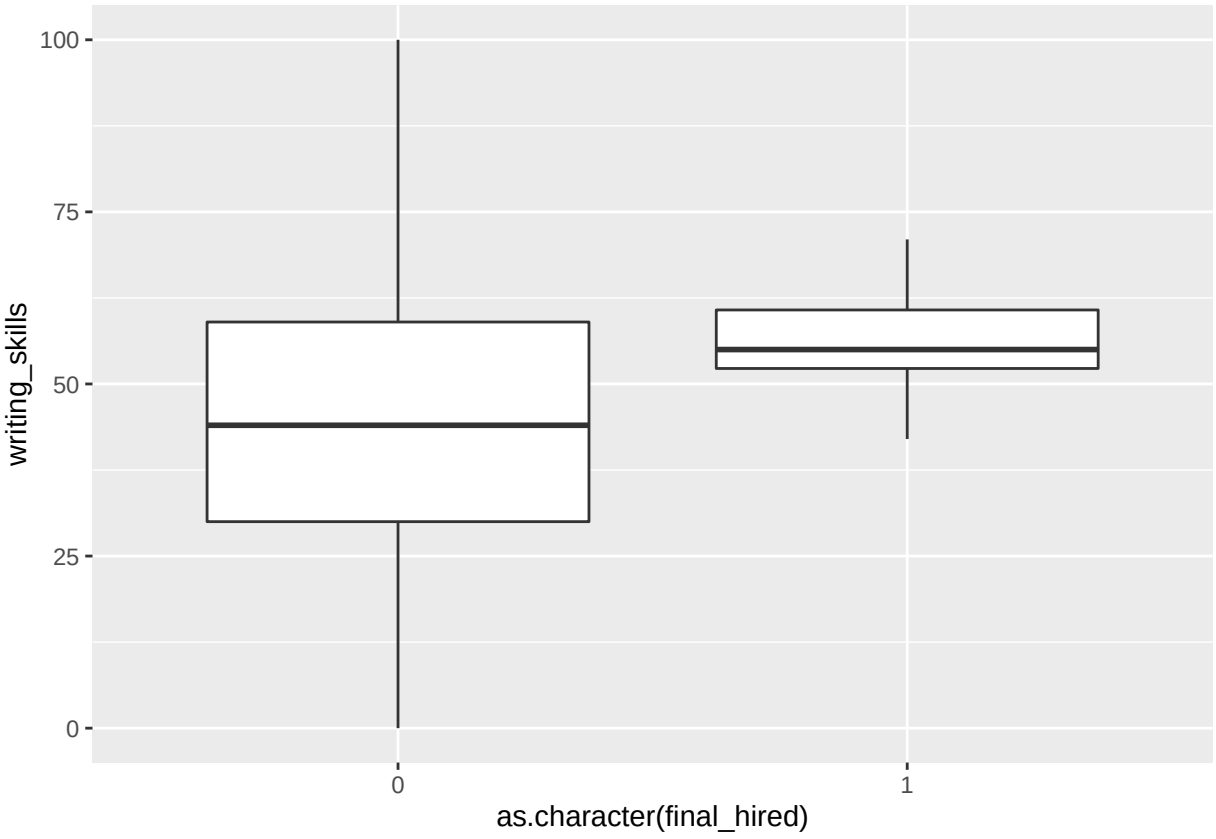


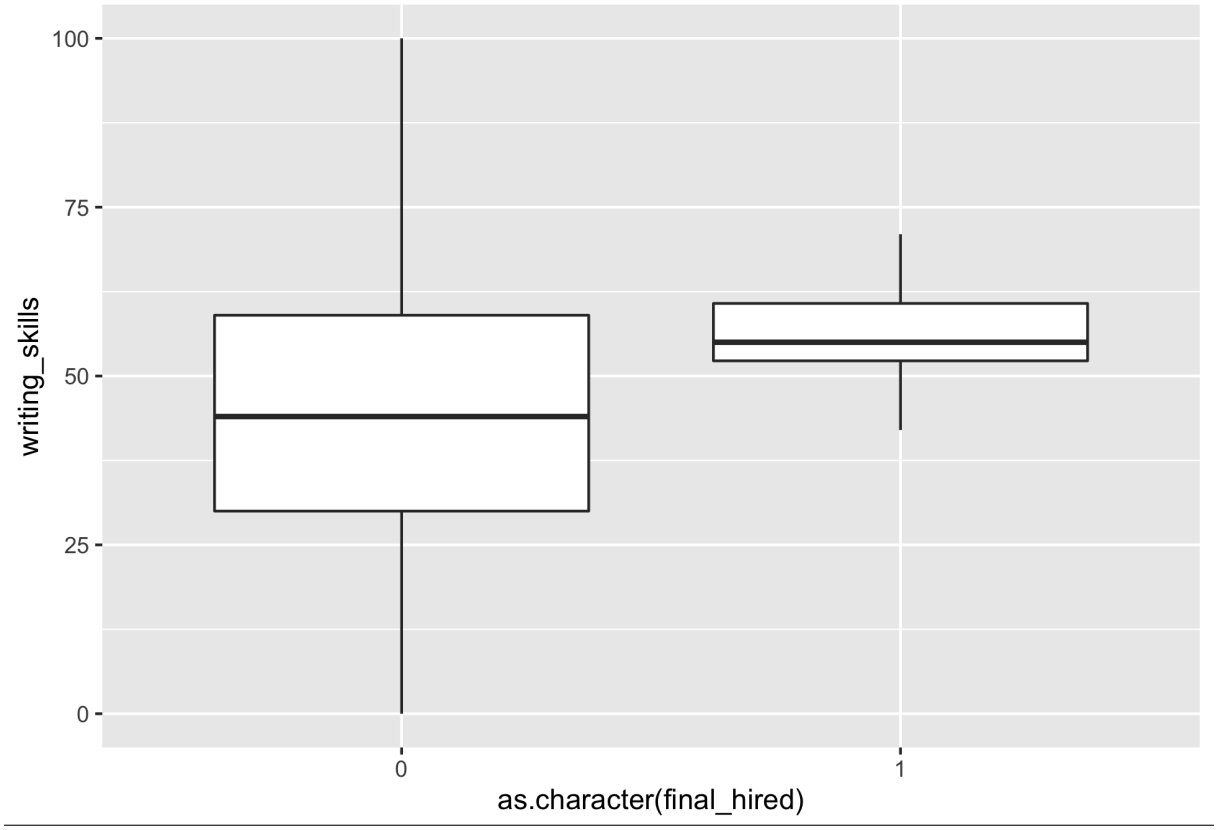

10.0
7.5
5.0
2.5
0.0
speaking_skills
0
1
as.character(pass_phase2)

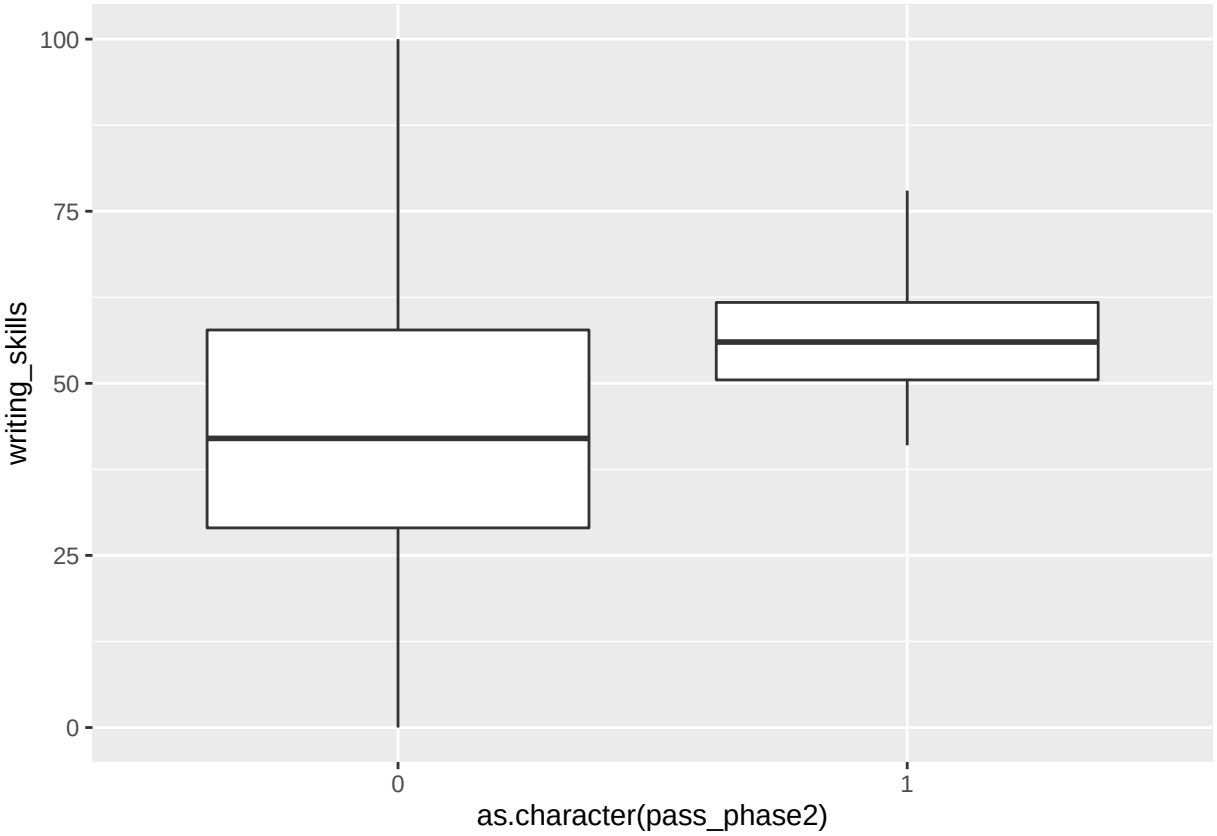

10.0
7.5
5.0
2.5
0.0
speaking_skills
0
1
as.character(final_hired)

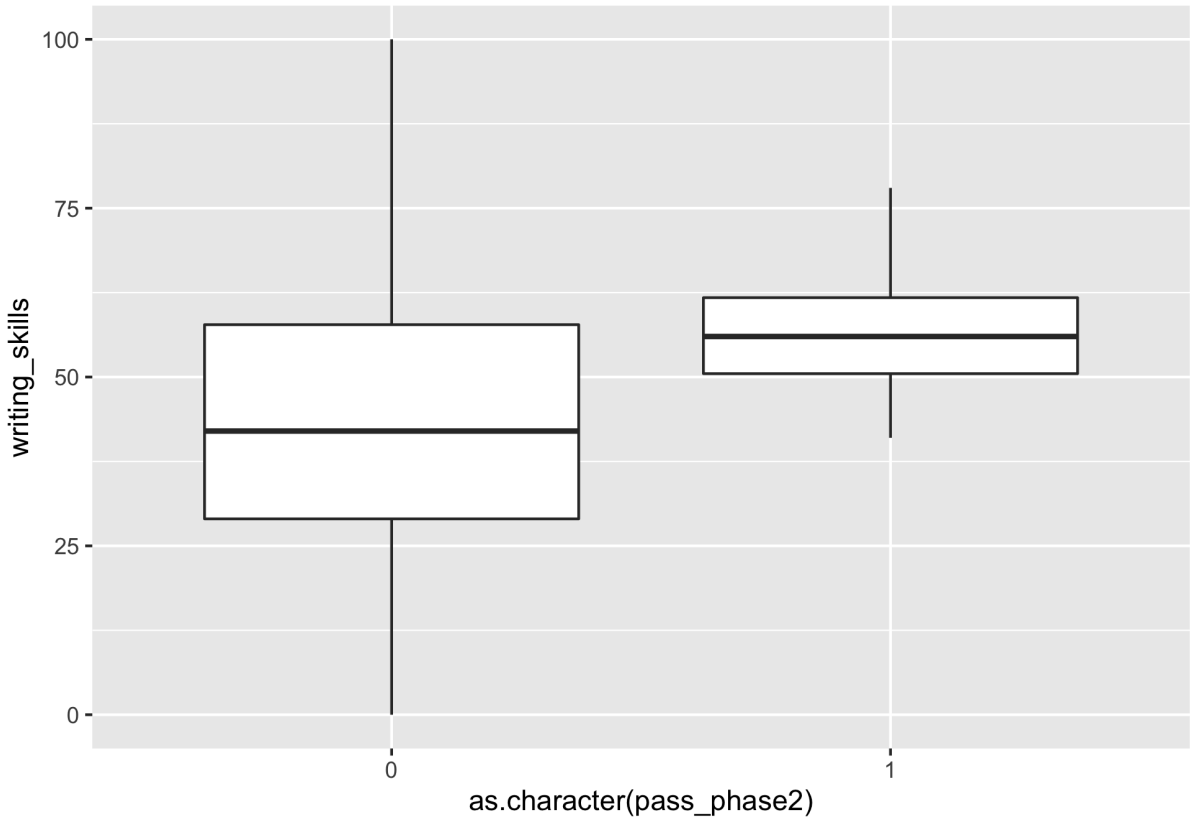

10.0
7.5
5.0
2.5
0.0
speaking_skills
0
1
as.character(pass_phase2)

Writing Skills VS if hired

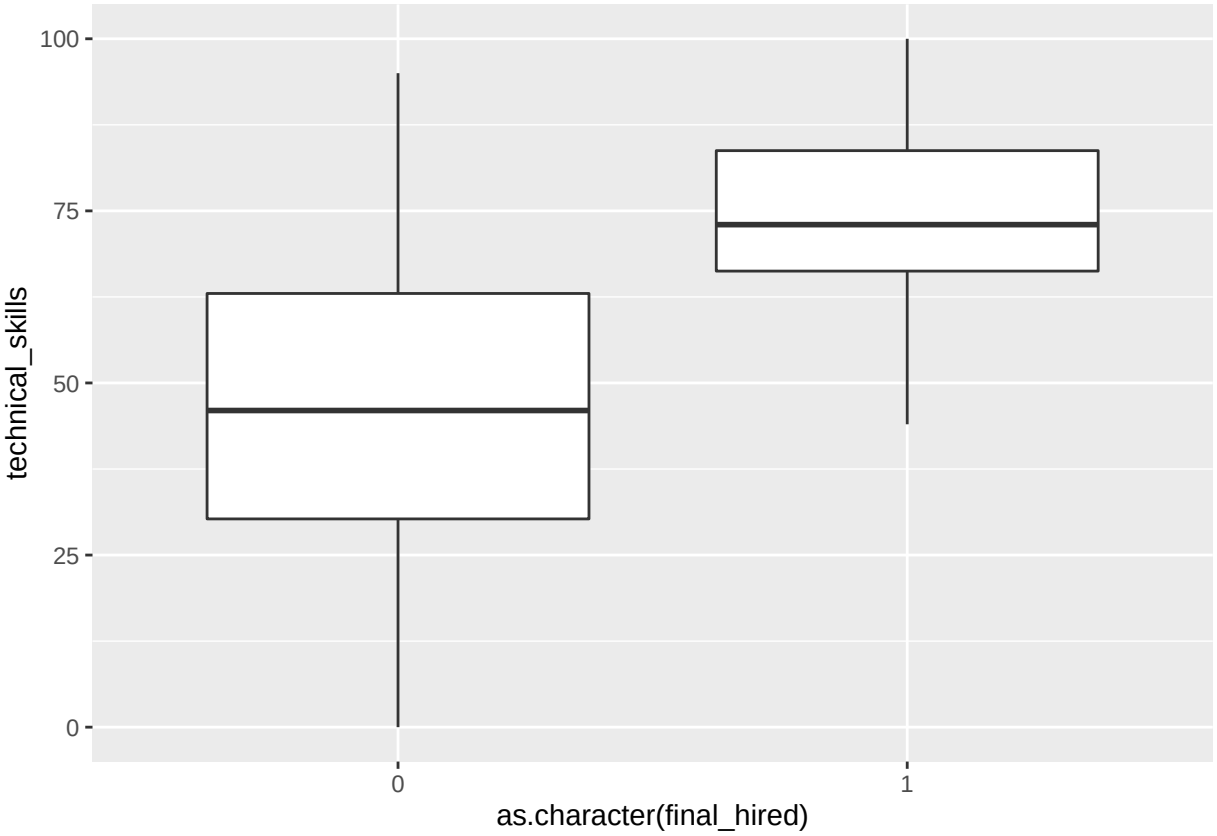


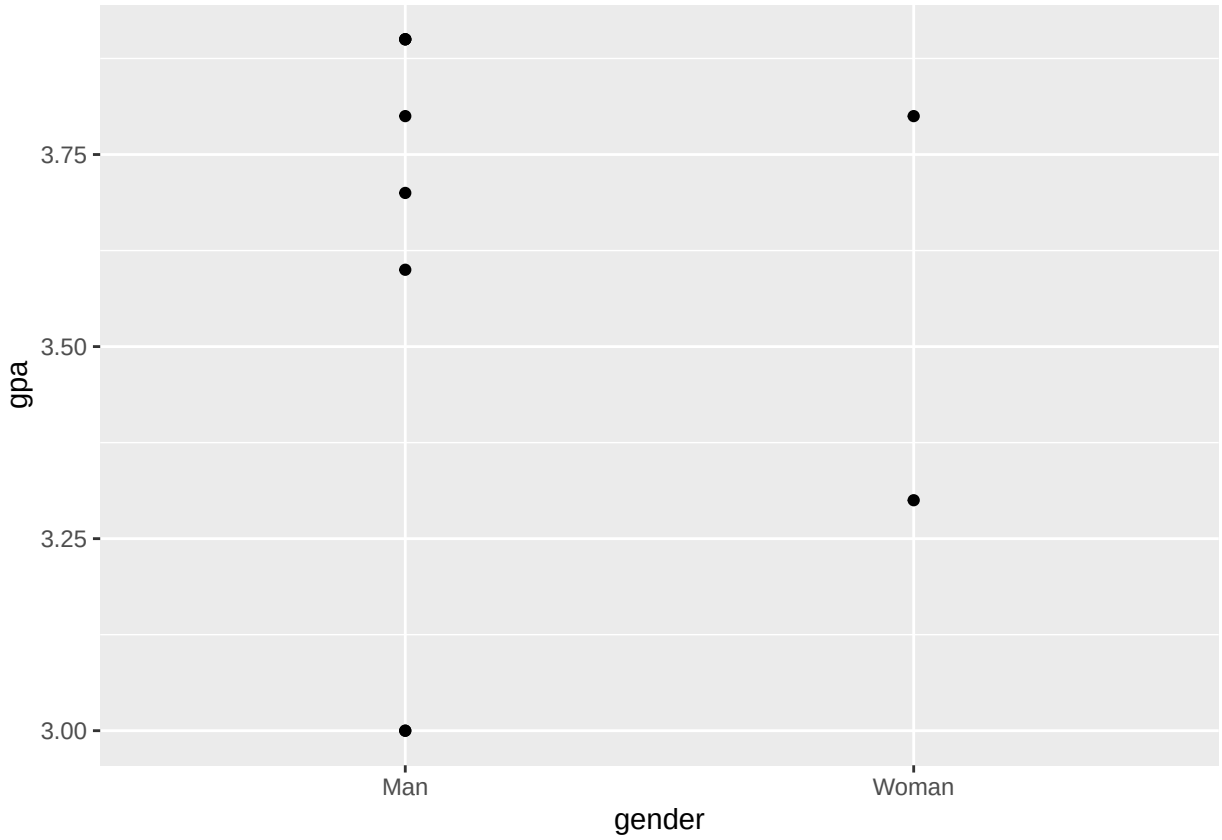


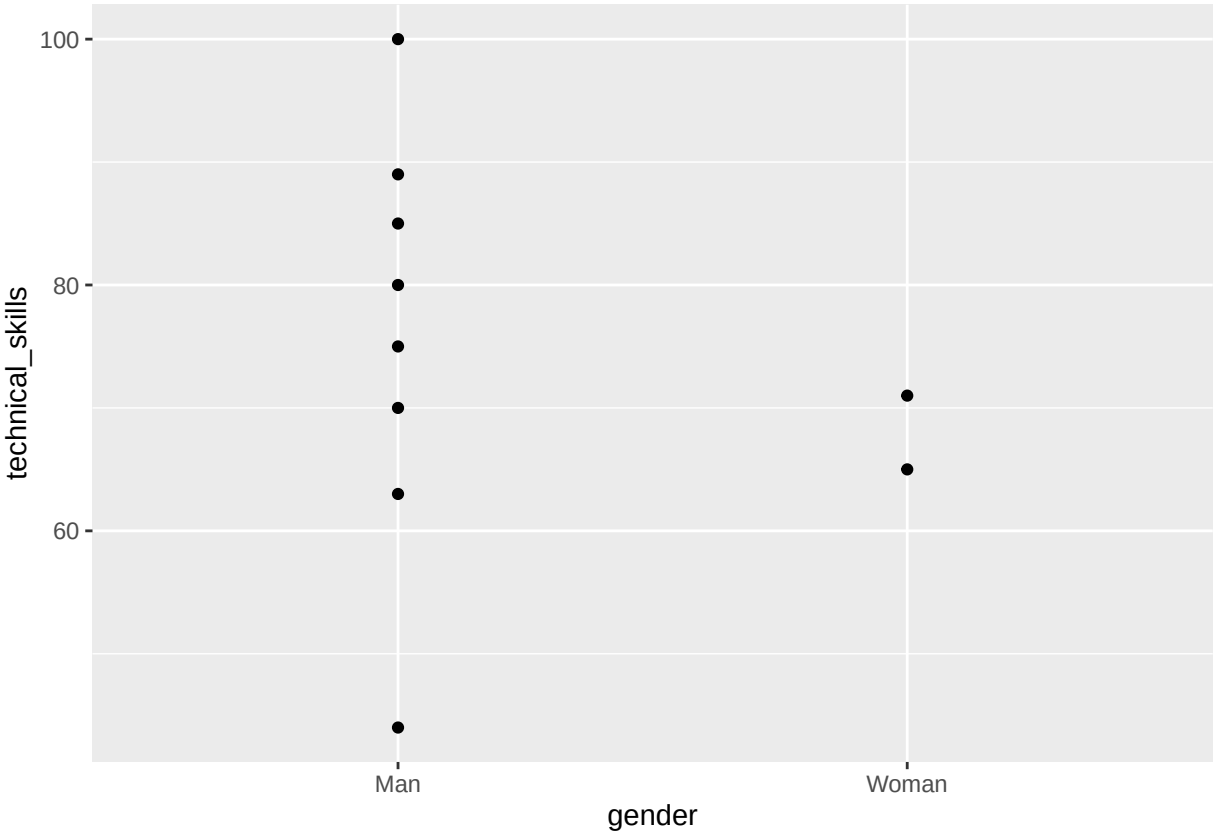

100
75
50
25
0
writing_skills
0
1
as.character(pass_phase2)

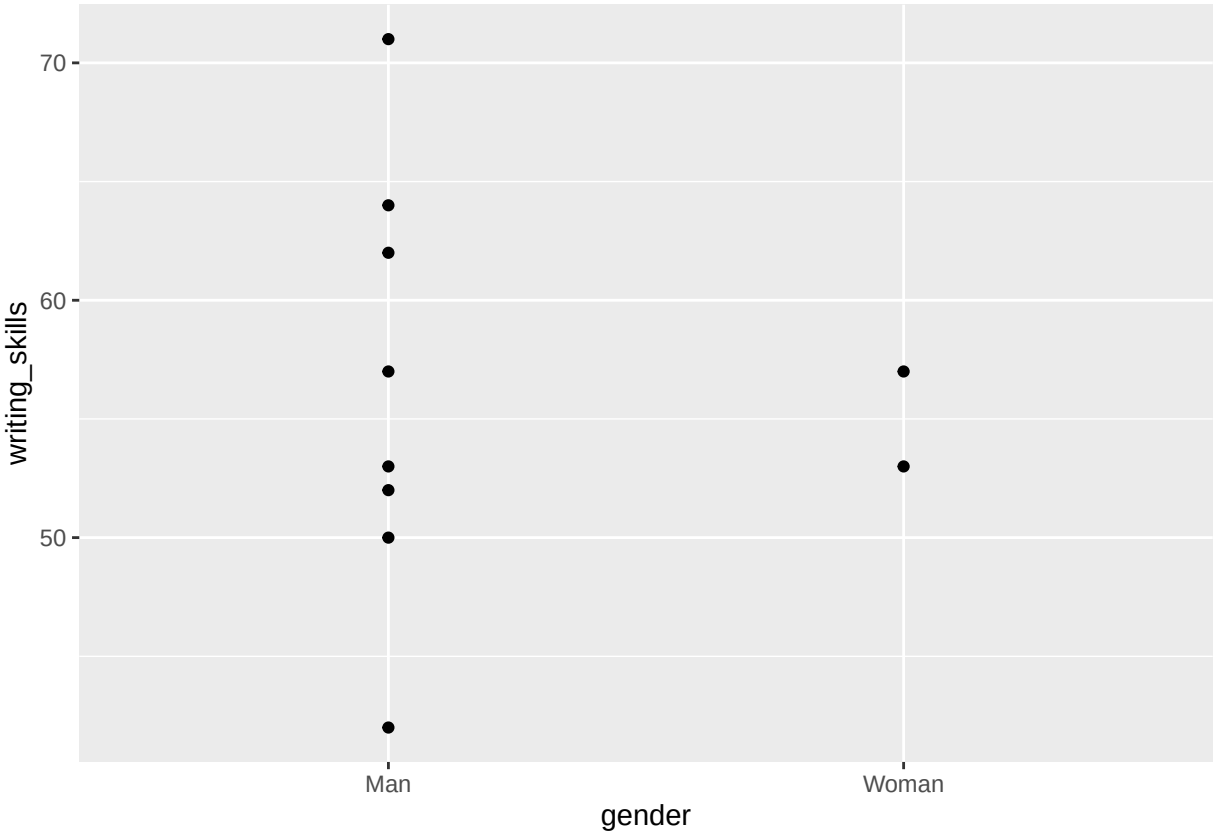

100
75
50
25
0
writing_skills
0
1
as.character(pass_phase2)

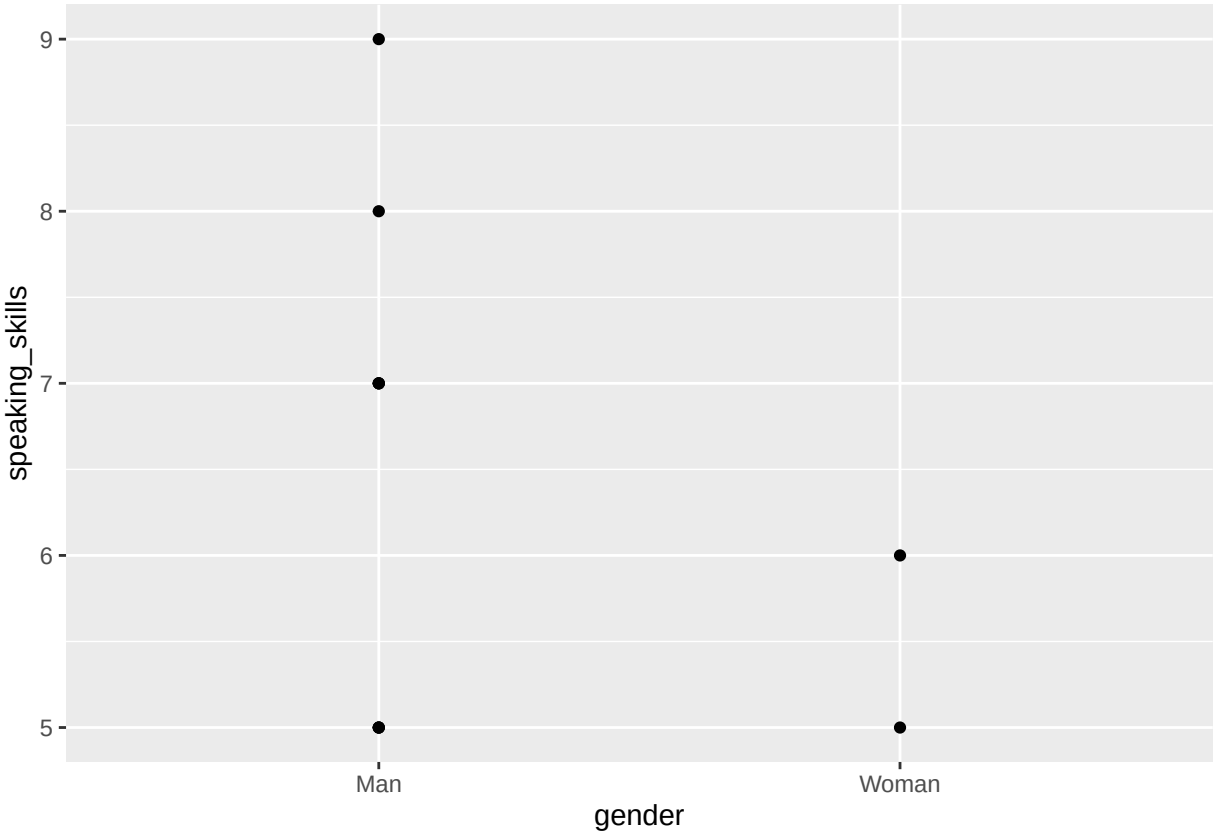
Tech Skills VS if hired

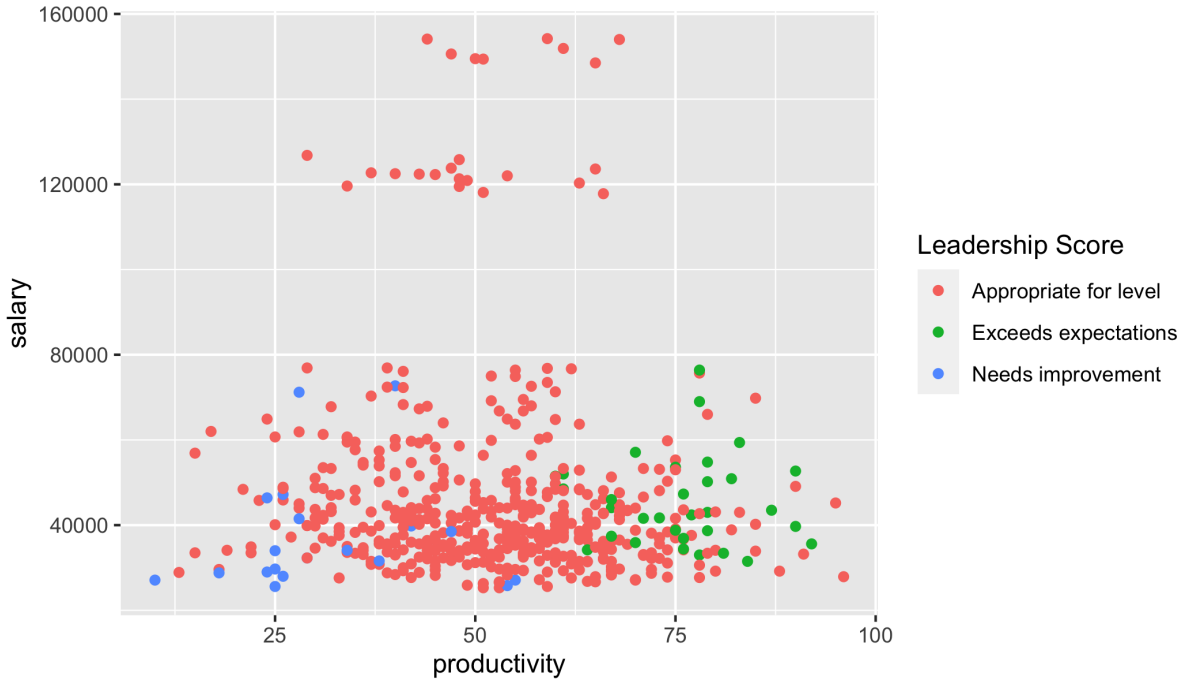


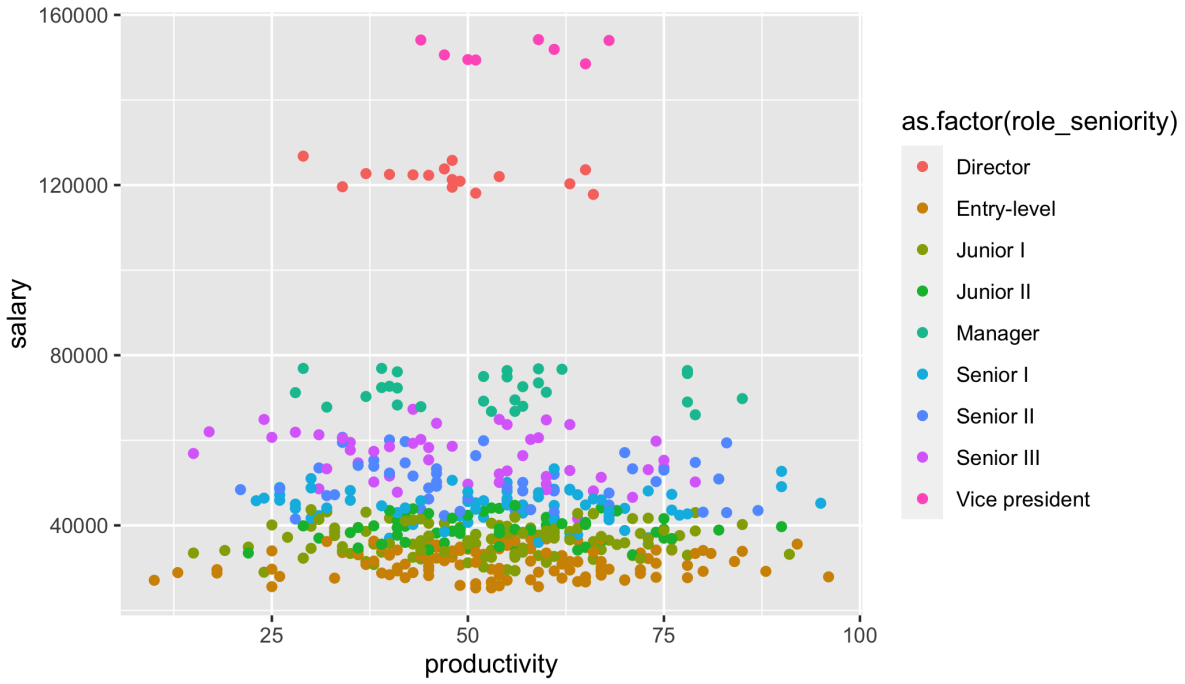

3.75
3.50
3.25
3.00
gpa
Man
Woman
gender


100
80
60
technical_skills
Man
Woman
gender


70
60
50
writing_skills
Man
Woman
gender


9
8
7
6
5
speaking_skills
Man
Woman
gender


160000
120000
80000
40000
salary
25
50
75
100
productivity
Leadership Score
Appropriate for level
Exceeds expectations
Needs improvement




```
##
## Call:
## glm(formula = pass_phase1 ~ gender + gpa + extracurriculars +
##     cv + work_experience, family = binomial(link = "logit"),
##     data = phase1_new_applicants)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.60450  -0.64746  -0.00004   0.68146   1.96684
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -25.16292  648.71016  -0.039  0.96906
## genderPrefer not to say    0.16339    0.85121   0.192  0.84778
## genderWoman               -0.05912    0.22001  -0.269  0.78815
## gpa                        2.09045    0.23547   8.878  < 2e-16 ***
## extracurriculars           0.28921    0.21330   1.356  0.17514
## cv                        18.68461  648.70981   0.029  0.97702
## work_experience            0.76135    0.27647   2.754  0.00589 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 849.52  on 612  degrees of freedom
## Residual deviance: 516.92  on 606  degrees of freedom
## AIC: 530.92
##
## Number of Fisher Scoring iterations: 17

##
## Call:
## glm(formula = pass_phase2 ~ gender + team_applied_for + cover_letter +
##     extracurriculars + work_experience + technical_skills + writing_skills +
##     leadership_presence + speaking_skills, family = binomial(link = "logit"),
##     data = phase2_new_applicants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7705  -0.1309  -0.0242  -0.0045   3.2873
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -24.15050    4.79613  -5.035 4.77e-07 ***
## genderPrefer not to say   -16.20043 1974.74800  -0.008   0.9935
## genderWoman                -0.63266    0.79481  -0.796   0.4260
## team_applied_forSoftware    1.40910    0.76203   1.849   0.0644 .
## cover_letter                     NA         NA      NA       NA
## extracurriculars           -0.63485    0.71598  -0.887   0.3752
## work_experience            -0.10831    0.73646  -0.147   0.8831
## technical_skills            0.09897    0.02490   3.974 7.06e-05 ***
## writing_skills              0.10690    0.02747   3.892 9.93e-05 ***
## leadership_presence         1.00449    0.22639   4.437 9.13e-06 ***
## speaking_skills             0.90524    0.21952   4.124 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 157.306  on 299  degrees of freedom
## Residual deviance:  64.515  on 290  degrees of freedom
## AIC: 84.515
##
## Number of Fisher Scoring iterations: 16

##
## Call:
## glm(formula = final_hired ~ gender, family = binomial(link = "logit"),
##     data = phase3_new_applicants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2346  -1.2346  -0.8203   1.1213   1.5829
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1335     0.5175   0.258    0.796
## genderWoman  -1.0498     0.9838  -1.067    0.286
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.316  on 21  degrees of freedom
## Residual deviance: 29.103  on 20  degrees of freedom
## AIC: 33.103
##
## Number of Fisher Scoring iterations: 4

##
##   61 61.5   66   68 70.5   72 72.5   74 74.5 75.5 76.5   77 77.5   78   80 81.5
##    1    1    1    1    1    1    1    1    1    2    2    1    1    2    1    1
## 83.5 84.5 90.5
##    1    1    1
```

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

- Dataset size tooooo small!! especially the final hired data and the phase 3 data (22 observations)

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.