Diversity

Tianpei Qian

2023-01-18

TL;DR

- Sign on bonus is given out equally across gender and degree level but very differently across departments. HR department has the lowest bonus rate while the engineering department has the highest.
- The HR department hire much more females than males while we observe the opposite in all the other departments.
- A female is less likely to be on the managing level compared with a male in the sample department, with the same degree level and same years of experience! The problem seems to even get worse at higher levels.
- Salary-wise, the engineering department has the highest salary while the HR department has the lowest. We don't see gender inequality in salary.

Step 1: read data

```
company_hierarchy_file <- 'company_hierarchy.csv'
employee_file <- 'employee.csv'

company_hierarchy <-
  read_csv(
    company_hierarchy_file,
    col_types = list(col_integer(), col_integer(), col_character())
  )
```

```
employee <-
  read_csv(
    employee_file,
    col_types =
      list(
        col_integer(),
        col_integer(),
        col_double(),
        col_character(),
        col_character(),
        col_double()
      )
  )
```

Step 2: data cleaning

Join tables, process factors, etc.

```
employee_combined <-
  employee %>%
  inner_join(company_hierarchy, by = c("employee_id" = "employee_id")) %>%
  mutate(
    sex = factor(sex, levels = c("M", "F")),
    degree_level = factor(degree_level, levels = c("High_School","Bachelor","Master","PhD")),
    dept = factor(dept, levels = c("engineering","HR","marketing","sales", "CEO"))
  )
```

Step 3: quick check

Mainly check on 1-dimensional data. Look out for outliers/missing values.

```
employee_combined %>%
  summary()
```

```
##   employee_id     signing_bonus        salary            degree_level  sex
##  Min.   :    40   Min.   :0.0000   Min.   : 60000   High_School:1657   M:6439
##  1st Qu.: 50574   1st Qu.:0.0000   1st Qu.:110000   Bachelor   :2735   F:3561
##  Median : 99244   Median :0.0000   Median :182000   Master     :2786
##  Mean   :100002   Mean   :0.3014   Mean   :189112   PhD        :2822
##  3rd Qu.:149748   3rd Qu.:1.0000   3rd Qu.:255000
##  Max.   :199956   Max.   :1.0000   Max.   :700000
##
##  yrs_experience      boss_id               dept
##  Min.   : 1.000   Min.   :    79   engineering:2696
##  1st Qu.: 2.000   1st Qu.: 55883   HR         :1694
##  Median : 3.000   Median :102712   marketing  :2010
##  Mean   : 3.875   Mean   :103300   sales      :3599
##  3rd Qu.: 5.000   3rd Qu.:152288   CEO        :   1
##  Max.   :34.000   Max.   :199950
##                   NA's   :1
```

First, CEO is a special department with only 1 data point. It probably also contributes the only missing data for boss_id.

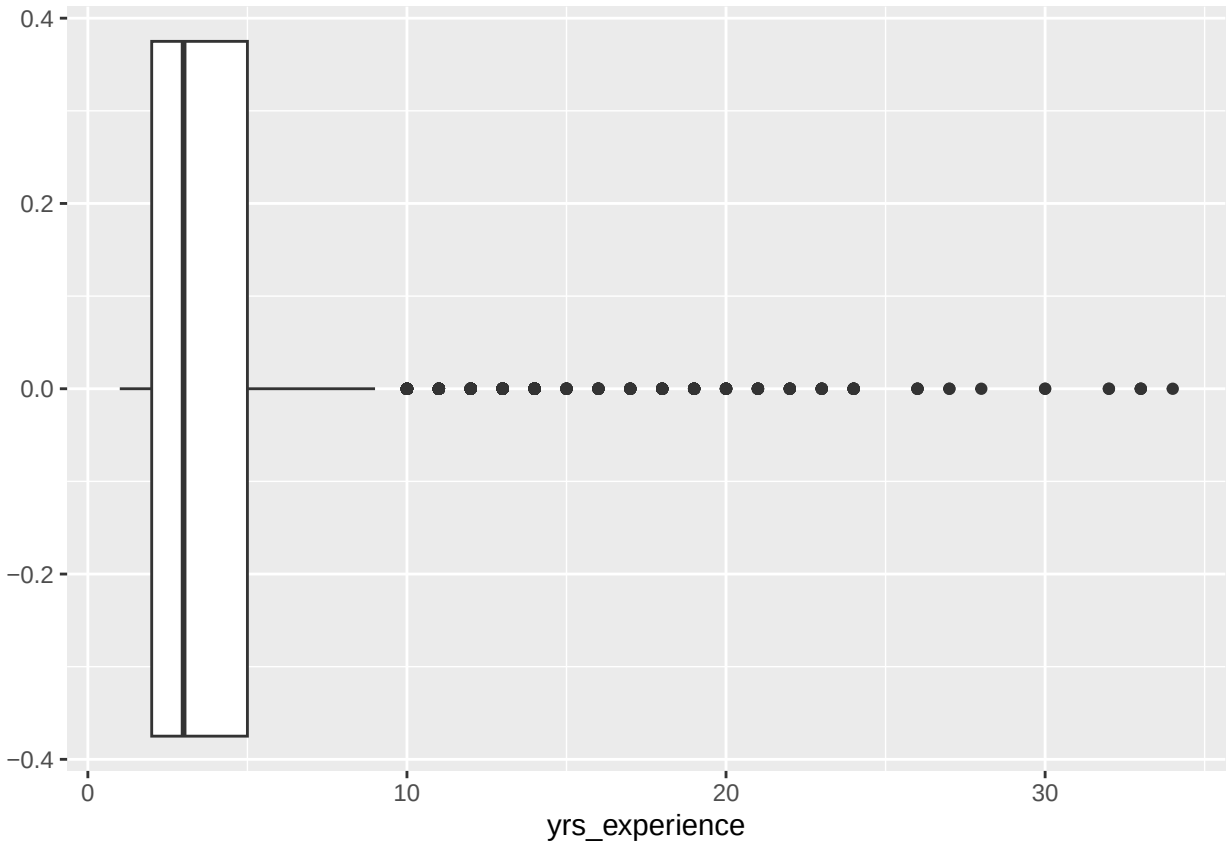
```
employee_combined %>%
  filter(dept == 'CEO')
```

```
## # A tibble: 1 x 8
##   employee_id signing_bonus salary degree_level sex   yrs_experi~1 boss_id dept
##         <int>         <int>  <dbl> <fct>        <fct>        <dbl>   <int> <fct>
## 1       61554             1 700000 PhD          M                7      NA CEO
## # ... with abbreviated variable name 1: yrs_experience
```

We may or may not remove this CEO data point. It won't cause a big problem if we keep it and handle it with care.

It looks like years of experience may have outliers.

```
employee_combined %>%
  ggplot(aes(yrs_experience)) +
  geom_boxplot()
```



Most employees have fewer than 10 years of expereince in the company.

Step 4: data transformation

Now that we're familiar with the data, let's derive some additional variables. Specifically, let's figure out the level of each employee and how many people they manage.

```
rank_levels <- c("IC", "MM", "D", "VP", "E", "CEO")
```

The following is in fact a wrong approach. It assumes that the tree is complete while there is a director that doesn't manage anyone!

```
employee_ranks_processed <-
  employee_combined %>%
  left_join(employee_combined, by = c("employee_id" = "boss_id"), suffix = c("", "_sub")) %>%
  filter(is.na(employee_id_sub)) %>%


  transmute(
    employee_id,
    boss_id,
    num_reports = 0,
    rank = factor("IC", levels = rank_levels)
  )

employee_ranks <- employee_ranks_processed

for (rank_level in rank_levels[2:length(rank_levels)]) {
```

```
  employee_ranks_processed <-
    employee_combined %>%
    inner_join(
      employee_ranks_processed, by = c("employee_id" = "boss_id"), suffix = c("", "_sub")
    ) %>%
    group_by(employee_id) %>%
    summarize(
      boss_id = max(boss_id),
      num_reports = sum(num_reports) + n()
    ) %>%
    mutate(
      rank = factor(rank_level, levels = rank_levels)
    )
  employee_ranks <- employee_ranks %>% rbind(employee_ranks_processed)
}
```

To get the rank, we have to start from the root of the tree, i.e. CEO.

```
employee_ranks <-
  employee_combined %>%
  mutate(
    rank = if_else(is.na(boss_id), "CEO", "to be updated")
  )

for (i in 1:(length(rank_levels) - 1)) {
  employee_ranks <-
    employee_ranks %>%
    left_join(
      employee_ranks,
      by = c("boss_id" = "employee_id"),
      suffix = c("", "_boss")
    ) %>%
    mutate(
      rank =
        if_else(
          rank_boss == rank_levels[length(rank_levels) - i + 1],
          rank_levels[length(rank_levels) - i],
          rank,
          rank # be careful! rank_boss can be null
        )
    ) %>%
    select_at(vars(-contains("_boss")))
}
```

Next we need to start from the leaf of the tree to get the number of reports for each employee.

```
employee_reports <-
  employee_ranks %>%
  filter(rank == rank_levels[1]) %>%
  transmute(
    employee_id,
    boss_id,
    num_reports = 0
  )
```

```
for (rank_level in rank_levels[-1]) {
  employee_reports <-
    employee_ranks %>%
    filter(rank == rank_level) %>%
    left_join(employee_reports, by = c("employee_id" = "boss_id"), suffix = c("", "_sub")) %>%
    group_by(employee_id, boss_id) %>%
    summarize(
      num_reports = sum(num_reports) + n(),
      # suffix only comes into play when there's duplicate!!
      .groups = "drop"
    ) %>%
    mutate(num_reports = if_else(is.na(num_reports), 0, num_reports)) %>%
    rbind(employee_reports)
}
```

Combine everything and take a final look

```
employee_final <-
  employee_ranks %>%
  inner_join(employee_reports, by = c("employee_id" = "employee_id", "boss_id" = "boss_id")) %>%
  filter(rank != "CEO") %>%
  mutate(rank = factor(rank, levels = rank_levels))

employee_final %>% summary()
```

```
##   employee_id     signing_bonus        salary            degree_level  sex
##  Min.   :    40   Min.   :0.0000   Min.   : 60000   High_School:1657   M:6438
##  1st Qu.: 50573   1st Qu.:0.0000   1st Qu.:110000   Bachelor   :2735   F:3561
##  Median : 99258   Median :0.0000   Median :182000   Master     :2786
##  Mean   :100006   Mean   :0.3013   Mean   :189061   PhD        :2821
##  3rd Qu.:149752   3rd Qu.:1.0000   3rd Qu.:255000
##  Max.   :199956   Max.   :1.0000   Max.   :650000
##  yrs_experience      boss_id               dept       rank
##  Min.   : 1.000   Min.   :    79   engineering:2696   IC :9000
##  1st Qu.: 2.000   1st Qu.: 55883   HR         :1694   MM : 800
##  Median : 3.000   Median :102712   marketing  :2010   D  : 160
##  Mean   : 3.875   Mean   :103300   sales      :3599   VP :  35
##  3rd Qu.: 5.000   3rd Qu.:152288   CEO        :   0   E  :   4
##  Max.   :34.000   Max.   :199950                      CEO:   0
##   num_reports
##  Min.   :   0.000
##  1st Qu.:   0.000
##  Median :   0.000
##  Mean   :   3.876
##  3rd Qu.:   0.000
##  Max.   :3598.000
```

Another approach is to use recursion, which is capcable of traversing the tree top-down and bottom-up at one go.

```
employee_combined_re <-
  employee_combined %>%
  mutate(
    boss_id = if_else(is.na(boss_id), -1L, boss_id),
    rank = as.character(NA),
    num_reports = as.integer(NA)
```

```
  )

find_rank <- function(rank_idx, id) {
  employee_combined_re$rank[employee_combined_re$employee_id == id] <<- rank_levels[rank_idx]

  if (rank_idx == 1) {
    employee_combined_re$num_reports[employee_combined_re$employee_id == id] <<- 0
    return(0)
  }

  ct <- 0
  for (id_sub in employee_combined_re$employee_id[employee_combined_re$boss_id == id] ) {
    ct <- ct + find_rank(rank_idx - 1, id_sub) + 1
  }
  employee_combined_re$num_reports[employee_combined_re$employee_id == id] <<- ct
  return(ct)
}

find_rank(
  length(rank_levels),
  employee_combined_re %>% filter(boss_id == -1) %>% pull(employee_id)
)
```

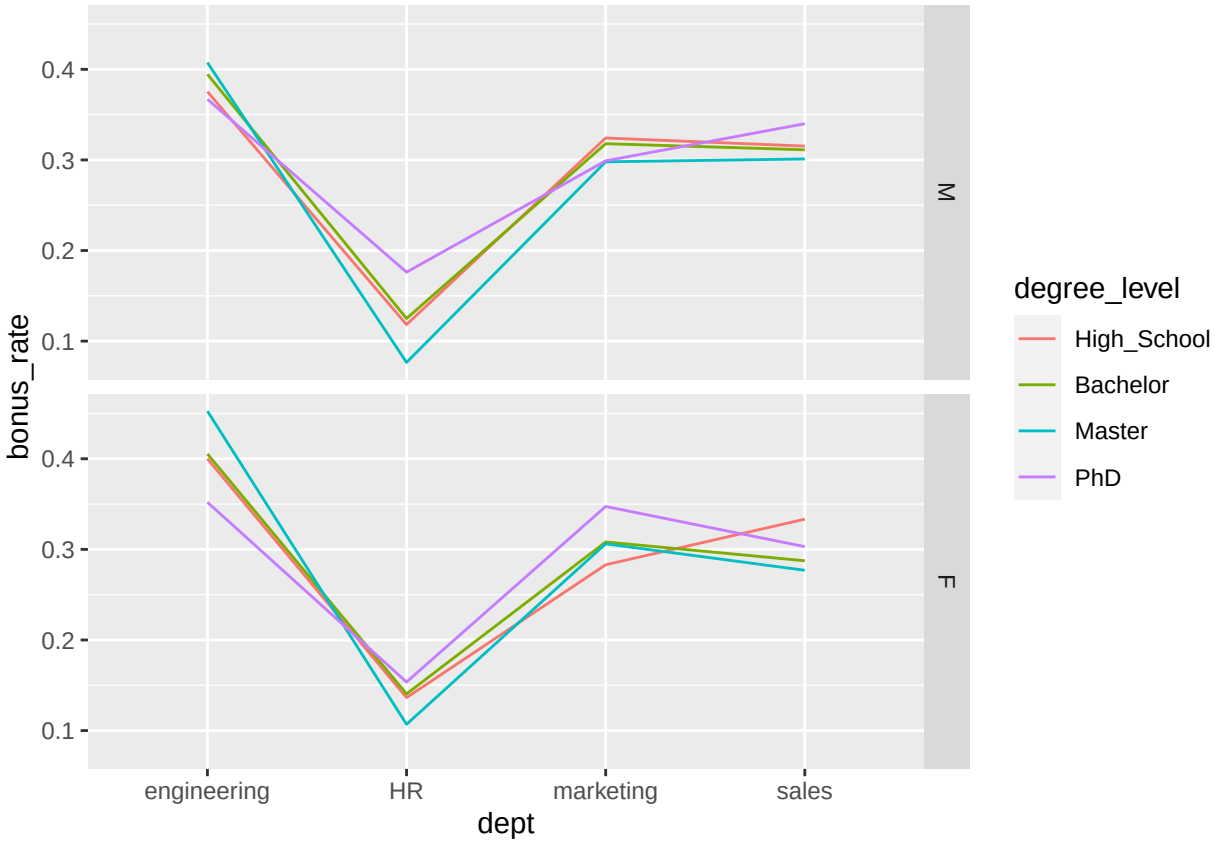
Step 4: exploratory analysis

Now we're finally ready to answer the million dollar question: do you think the company has been treating all its employees fairly?

Sign on bonus Is sign on bonus given out fairly?

```
employee_final %>%
  group_by(dept, degree_level, sex) %>%
  summarize(
    bonus_rate = sum(signing_bonus) / n()
  ) %>%
  ggplot(aes(dept, bonus_rate)) +
  geom_line(aes(color = degree_level, group = degree_level)) +
  facet_grid(vars(sex))
```

```
## `summarise()` has grouped output by 'dept', 'degree_level'. You can override
## using the `.groups` argument.
```



A quick plot shows that sex and degree level seems irrelevant to bonus rate but it seems to be lower in the HR department. There is no obvious interaction pattern between these variables.

```
bonus_lm <-
  glm(
    signing_bonus ~ degree_level + sex + dept,
    data = employee_final,
    family = 'binomial'
  )
```

```
bonus_lm %>% summary()
```

```
##
## Call:
## glm(formula = signing_bonus ~ degree_level + sex + dept, family = "binomial",
##     data = employee_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0045  -0.8696  -0.8501   1.3691   2.0326
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.438421   0.064123  -6.837 8.08e-12 ***
## degree_levelBachelor -0.001589   0.068911  -0.023    0.982
## degree_levelMaster   -0.038613   0.068821  -0.561    0.575
## degree_levelPhD       0.017177   0.068357   0.251    0.802
```

```
## sexF                -0.002431   0.048297  -0.050    0.960
## deptHR              -1.450644   0.084214 -17.226  < 2e-16 ***
## deptmarketing       -0.354766   0.062449  -5.681 1.34e-08 ***
## deptsales           -0.354036   0.053597  -6.606 3.96e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12239  on 9998  degrees of freedom
## Residual deviance: 11864  on 9991  degrees of freedom
## AIC: 11880
##
## Number of Fisher Scoring iterations: 4
```

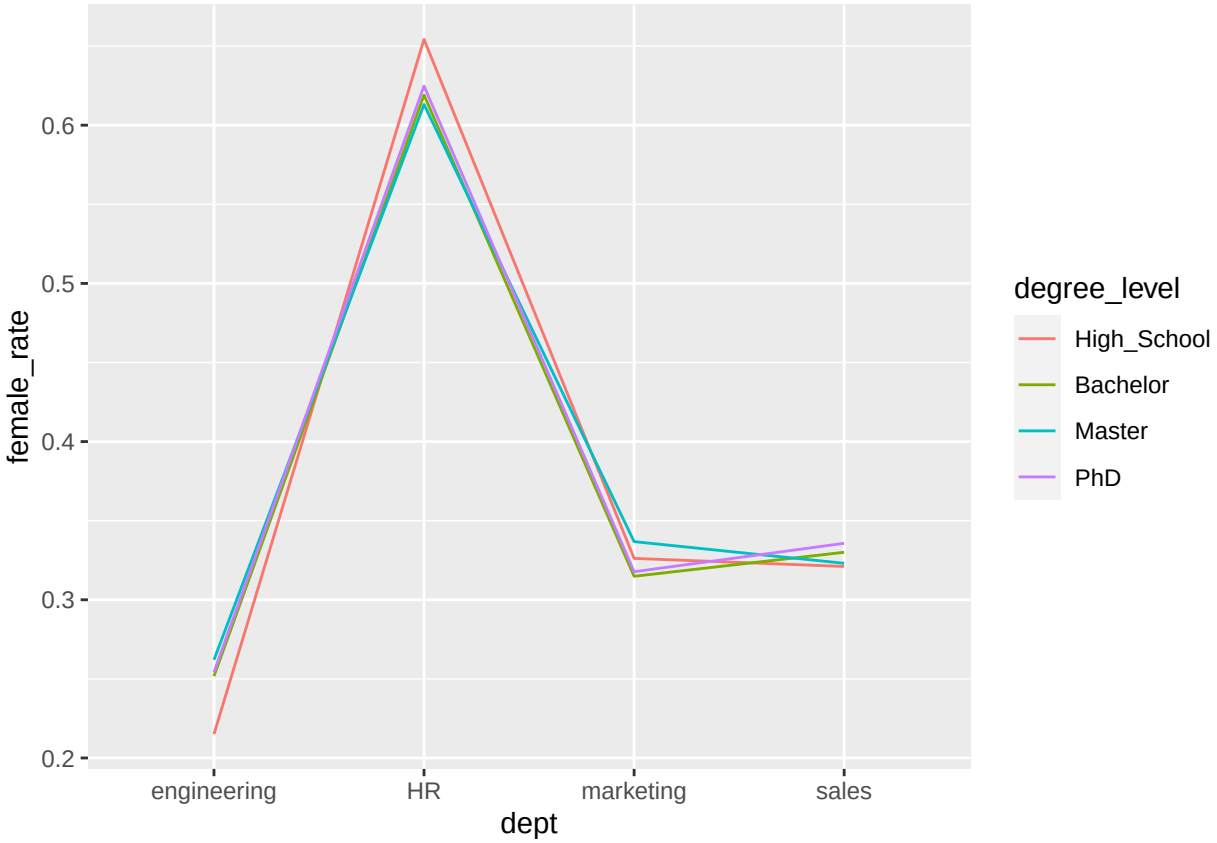
Indeed, department has a significant impact on bonus rates, with the HR department being the lowest and the engineering department being the highest.

Note that in answering this question we shouldn't consider more variables than needed!! This is a common pitfall that people new to modelling can fall into. For example, gender may affect the level the person gets and the level can subsequently decide whether the person gets sign on bonus. If you include "level" in your model when studying the impact of gender, then you will get an inaccurate estimate because you limit the impact of gender on level.

Department Does each department have the same criteria for hiring?

```
employee_final %>%
  group_by(dept, degree_level) %>%
  summarize(
    female_rate = sum(sex == 'F') / n()
  ) %>%
  ggplot(aes(dept, female_rate)) +
  geom_line(aes(color = degree_level, group = degree_level))
```

```
## `summarise()` has grouped output by 'dept'. You can override using the
## `.groups` argument.
```



The HR department hires more females than males while the other 3 departments hire more males.

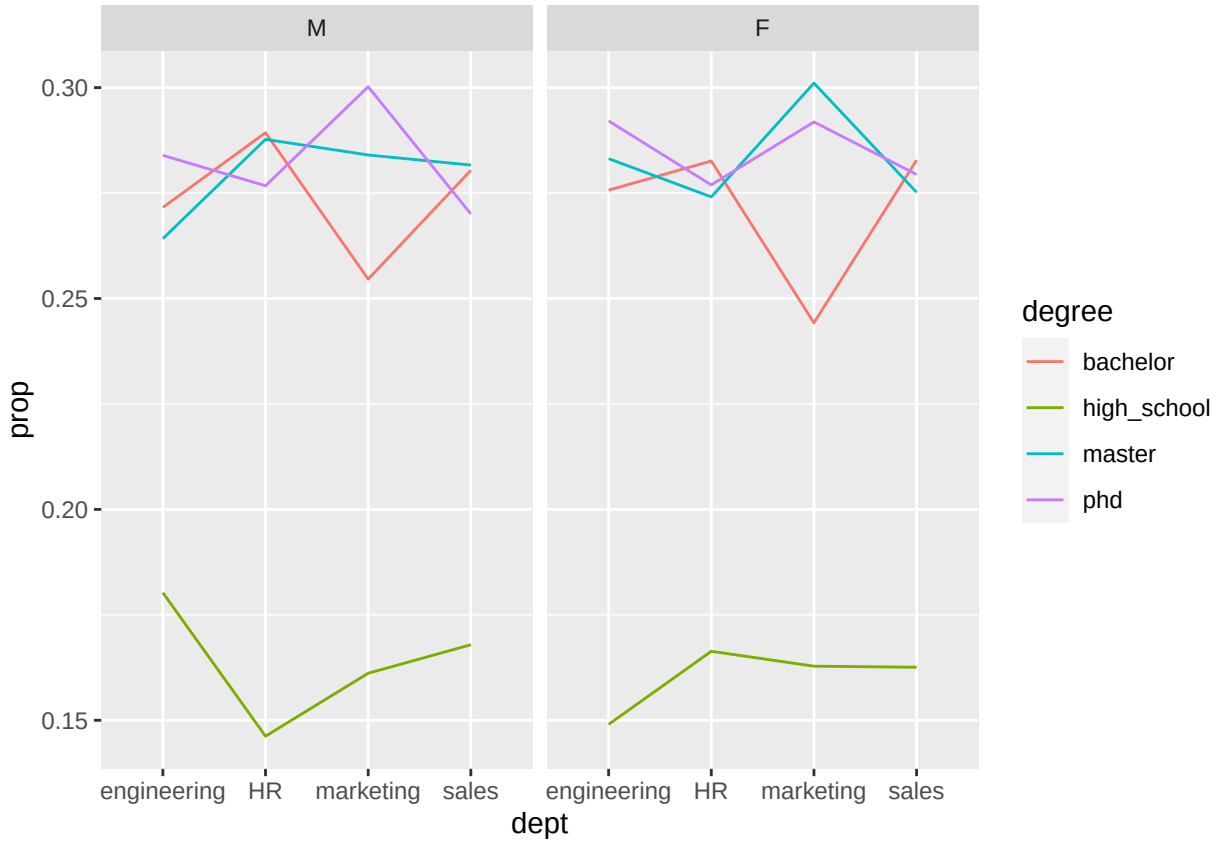
A chi-square test confirms that each department has significantly different gender distributions.

```
employee_final %>%
  count(sex, dept) %>%
  pivot_wider(names_from = sex, values_from = n) %>%
  select(-dept) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 688.93, df = 3, p-value < 2.2e-16
```

```
employee_final %>%
  group_by(dept, sex) %>%
  summarize(
    high_school = sum(degree_level == 'High_School') / n(),
    bachelor = sum(degree_level == 'Bachelor') / n(),
    master = sum(degree_level == 'Master') / n(),
    phd = sum(degree_level == 'PhD') / n()
  ) %>%
  pivot_longer(cols = high_school:phd, names_to = "degree", values_to = "prop") %>%
  ggplot() +
  geom_line(aes(dept, prop, color = degree, group = degree)) +
  facet_grid(cols = vars(sex))
```

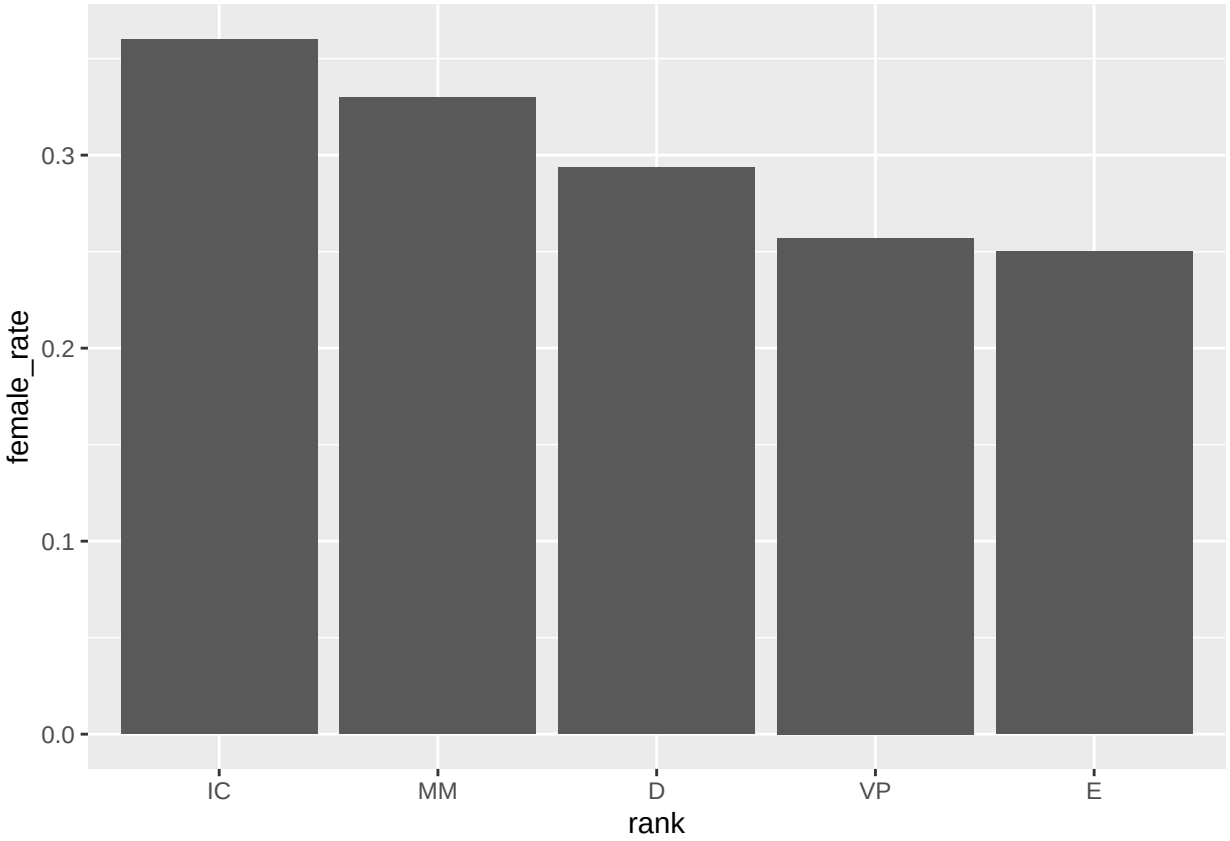
```
## `summarise()` has grouped output by 'dept'. You can override using the
## `.groups` argument.
```



On the other hand, all departments seem to have very similar hiring standards on degree levels.

Company level Does female have a fair representation on the managing levels?

```
employee_final %>%
  group_by(rank) %>%
  summarize(
    female_rate = sum(sex == "F") / n()
  ) %>%
  ggplot(aes(rank, female_rate)) +
  geom_col()
```



Firstly, the company has fewer females across all levels and female representation seems to get worse as the level increases.

To get a more rigorous analysis, let's build a model that also controls department, degree level and years of experience.

```
manager_lm <-
  glm(
    is_manager ~ degree_level + sex + dept + yrs_experience,
    data = employee_final %>% mutate(is_manager = (rank != "IC")),
    family = 'binomial'
  )
```

```
manager_lm %>% summary()
```

```
##
## Call:
## glm(formula = is_manager ~ degree_level + sex + dept + yrs_experience,
##     family = "binomial", data = employee_final %>% mutate(is_manager = (rank !=
##         "IC")))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3871  -0.3134  -0.1973  -0.1248   3.2789
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -6.202908   0.195528 -31.724  < 2e-16 ***
```

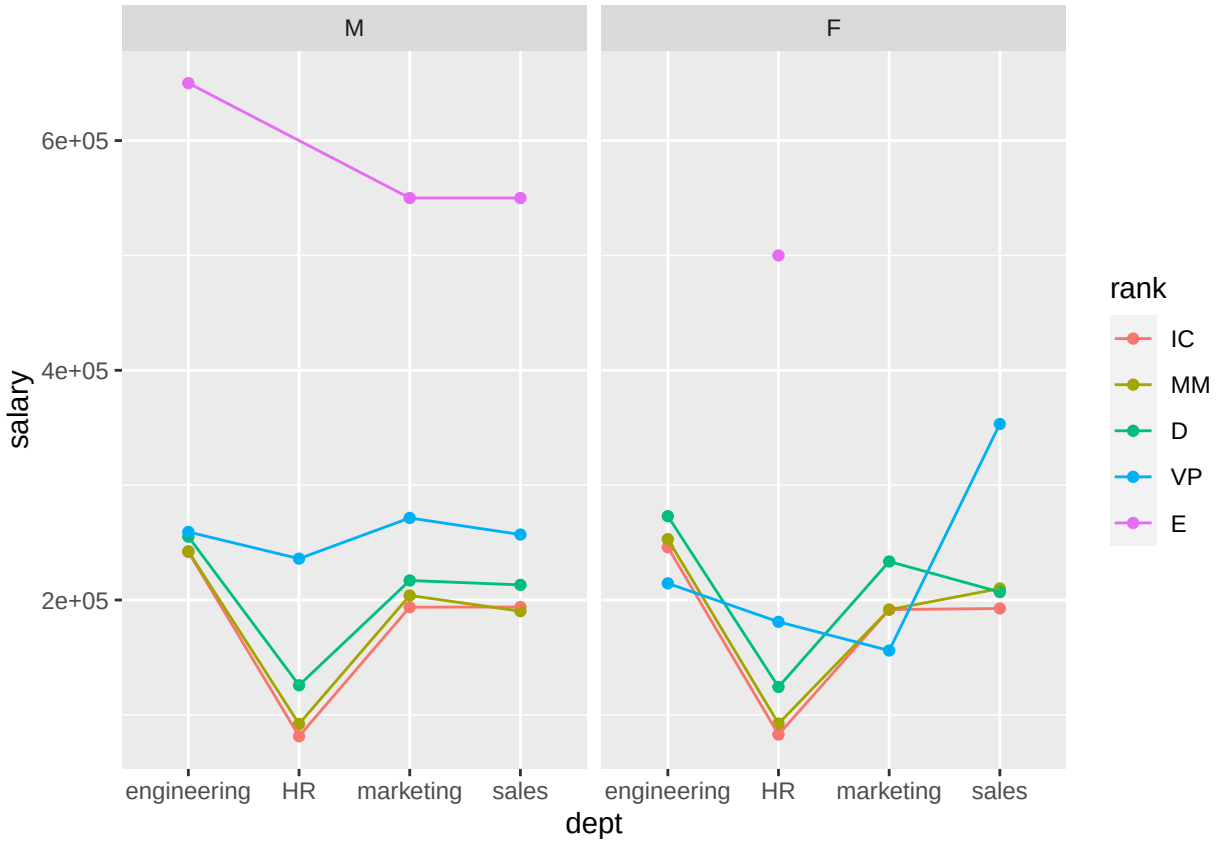
```
## degree_levelBachelor  0.592814   0.168647   3.515  0.00044 ***
## degree_levelMaster    0.987644   0.163942   6.024 1.70e-09 ***
## degree_levelPhD       1.029141   0.162856   6.319 2.63e-10 ***
## sexF                 -0.408918   0.096979  -4.217 2.48e-05 ***
## deptHR                0.111515   0.137917   0.809  0.41876
## deptmarketing         0.003487   0.130727   0.027  0.97872
## deptsales             0.049767   0.112082   0.444  0.65702
## yrs_experience        0.620473   0.017376  35.708  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6497.1  on 9998  degrees of freedom
## Residual deviance: 3776.0  on 9990  degrees of freedom
## AIC: 3794
##
## Number of Fisher Scoring iterations: 6
```

Not surprisingly, you have roughly equal opportunities to become managers in all departments. Also, higher degrees and more years of experience increase the odds. Lastly and notably, females are less likely to become managers in this company even when they're in the same department, have the same degree level and have the same years of experience!

Salary Finally, do we see inequality in salary?

```
employee_final %>%
  group_by(dept, rank, sex) %>%
  summarize(salary = mean(salary)) %>%
  ggplot(aes(dept, salary)) +
  geom_point(aes(color = rank, group = rank)) +
  geom_line(aes(color = rank, group = rank)) +
  facet_grid(cols = vars(sex))
```

```
## `summarise()` has grouped output by 'dept', 'rank'. You can override using the
## `.groups` argument.
```




```
salary_lm <-
  lm(
    salary ~ degree_level + sex + rank + dept  + yrs_experience + num_reports,
    data = employee_final
  )
```

```
salary_lm %>% summary()
```

```
##
## Call:
## lm(formula = salary ~ degree_level + sex + rank + dept + yrs_experience +
##     num_reports, data = employee_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -209583  -46463   -3156   45572  205822
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           239114.81    2365.77 101.073  < 2e-16 ***
## degree_levelBachelor    1702.67    2234.55   0.762  0.44609
## degree_levelMaster       806.01    2231.92   0.361  0.71801
## degree_levelPhD         2950.84    2229.69   1.323  0.18572
## sexF                     705.97    1554.07   0.454  0.64964
## rankMM                  2036.89    3046.85   0.669  0.50381
## rankD                  16608.56    7148.06   2.324  0.02017 *
## rankVP                 50212.35   18072.13   2.778  0.00547 **
```

```
## rankE                 279411.96  113395.20   2.464  0.01375 *
## deptHR               -159331.20    2298.90 -69.307  < 2e-16 ***
## deptmarketing         -49105.69    2117.19 -23.194  < 2e-16 ***
## deptsales             -49148.96    1831.06 -26.842  < 2e-16 ***
## yrs_experience           481.30     312.88   1.538  0.12401
## num_reports               37.54      42.96   0.874  0.38226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71710 on 9985 degrees of freedom
## Multiple R-squared:  0.3492, Adjusted R-squared:  0.3483
## F-statistic: 412.1 on 13 and 9985 DF,  p-value: < 2.2e-16
```

The department and company level matters for salary, while degree level, sex, years of experience and number of reports do not have a significant impact on salary for a given department and company level.