HR Analysis

Install all libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## Warning: package 'forcats' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(dplyr)
library(ggplot2)
library(lubridate)
library(tidyverse)
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.3.2
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(readr)
```

```
                  Step 1: Collect Data
```

Load both the train & test datasets

```
test <- read.csv("C:/Users/prane/OneDrive/Desktop/personal/Projects/HR Analysis/test.csv", he
ader=FALSE, stringsAsFactors=TRUE)

train <- read.csv("C:/Users/prane/OneDrive/Desktop/personal/Projects/HR Analysis/train.csv",
header=FALSE, stringsAsFactors=TRUE)
```

Understand the structure of the train data

```
head(train)
```

```
##            V1                V2        V3               V4     V5
## 1 employee_id        department    region        education gender
## 2       65438 Sales & Marketing  region_7 Master's & above      f
## 3       65141        Operations region_22       Bachelor's      m
## 4        7513 Sales & Marketing region_19       Bachelor's      m
## 5        2542 Sales & Marketing region_23       Bachelor's      m
## 6       48945        Technology region_26       Bachelor's      m
##                     V6              V7  V8                   V9
## 1 recruitment_channel no_of_trainings age previous_year_rating
## 2            sourcing               1  35                    5
## 3               other               1  30                    5
## 4            sourcing               1  34                    3
## 5               other               2  39                    1
## 6               other               1  45                    3
##                 V10           V11         V12                 V13         V14
## 1 length_of_service KPIs_met >80% awards_won? avg_training_score is_promoted
## 2                 8             1           0                  49           0
## 3                 4             0           0                  60           0
## 4                 7             0           0                  50           0
## 5                10             0           0                  50           0
## 6                 2             0           0                  73           0
```

```
colnames(test) <- unlist(test[1, ])
```

Remove the first row from the data frame

```
test <- test[-1, ]
```

Arrange the data frame by the employee_id column

```
head(test,2)
```

```
##   employee_id department    region  education gender recruitment_channel
## 2        8724 Technology region_26 Bachelor's      m            sourcing
## 3       74430         HR  region_4 Bachelor's      f               other
##   no_of_trainings age previous_year_rating length_of_service KPIs_met >80%
## 2               1  24                                     1             1
## 3               1  31                   3                 5             0
##   awards_won? avg_training_score
## 2           0                77
## 3           0                51
```

Understand the structure of the test data

```
head(train)
```

```
##            V1               V2        V3               V4     V5
## 1 employee_id        department    region        education gender
## 2       65438 Sales & Marketing  region_7 Master's & above      f
## 3       65141        Operations region_22       Bachelor's      m
## 4        7513 Sales & Marketing region_19       Bachelor's      m
## 5        2542 Sales & Marketing region_23       Bachelor's      m
## 6       48945        Technology region_26       Bachelor's      m
##                    V6             V7  V8                   V9
## 1 recruitment_channel no_of_trainings age previous_year_rating
## 2            sourcing               1  35                    5
## 3               other               1  30                    5
## 4            sourcing               1  34                    3
## 5               other               2  39                    1
## 6               other               1  45                    3
##                 V10           V11         V12                V13         V14
## 1 length_of_service KPIs_met >80% awards_won? avg_training_score is_promoted
## 2                 8             1           0                 49           0
## 3                 4             0           0                 60           0
## 4                 7             0           0                 50           0
## 5                10             0           0                 50           0
## 6                 2             0           0                 73           0
```

```
colnames(train) <- unlist(train[1, ])
```

Remove the first row (header row) from the data frame

```
train <- train[-1, ]
head(train,2)
```

```
##   employee_id        department    region        education gender
## 2       65438 Sales & Marketing  region_7 Master's & above      f
## 3       65141        Operations region_22       Bachelor's      m
##   recruitment_channel no_of_trainings age previous_year_rating
## 2            sourcing               1  35                    5
## 3               other               1  30                    5
##   length_of_service KPIs_met >80% awards_won? avg_training_score is_promoted
## 2                 8             1           0                 49           0
## 3                 4             0           0                 60           0
```

Arrange the data frame by the employee_id column

train %>% arrange(employee_id)

```
                      Step 2: Clean Data
```

Remove duplicates from train data

```
train <- distinct(train)
```

```
                   Step 3: Analyse and Visulaize Data
```

```
summary(train)
```

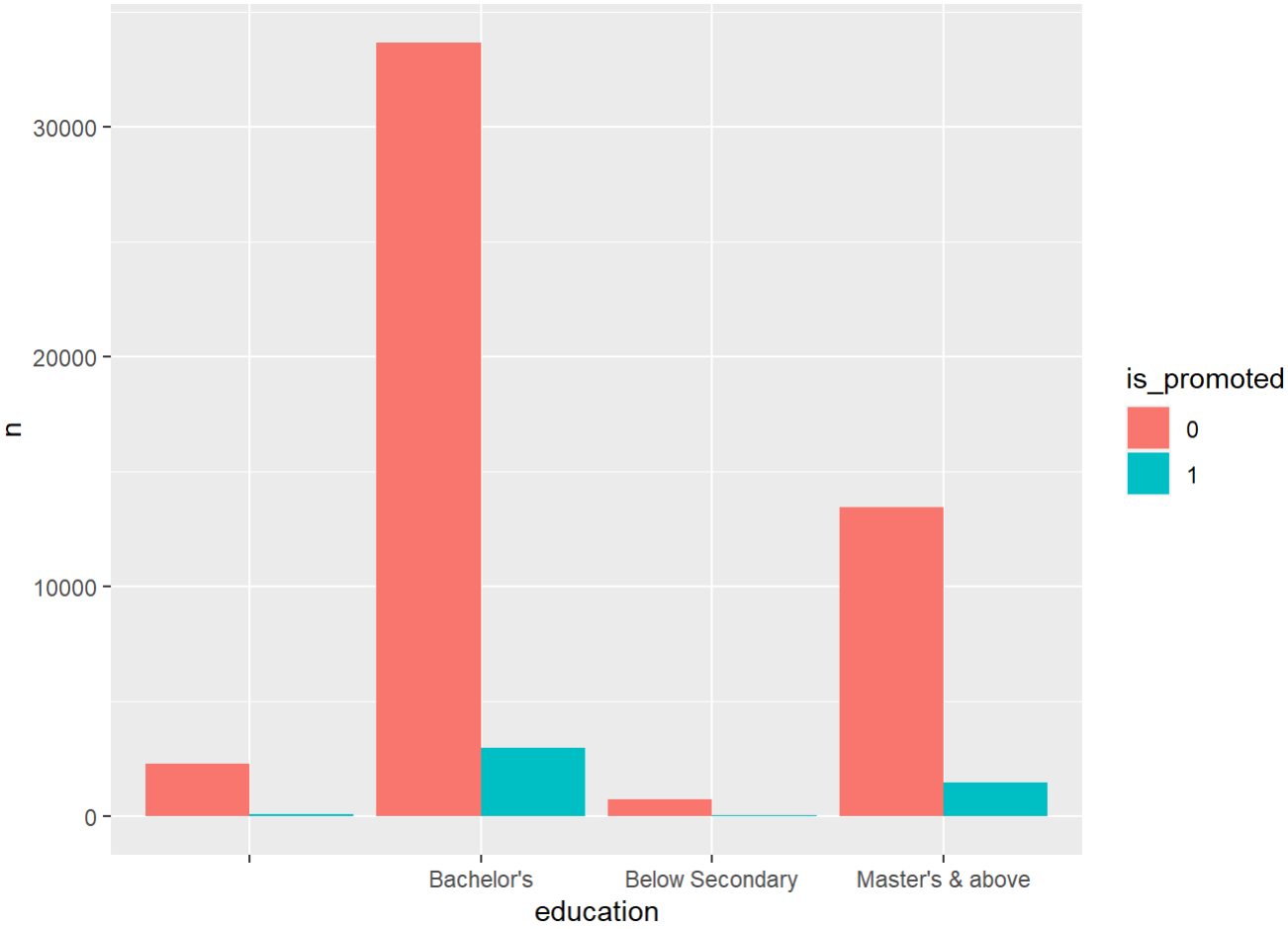
```
##   employee_id               department          region
##  1      :    1   Sales & Marketing:16840   region_2 :12343
##  10     :    1   Operations       :11348   region_22: 6428
##  100    :    1   Procurement      : 7138   region_7 : 4843
##  1000   :    1   Technology       : 7138   region_15: 2808
##  10000  :    1   Analytics        : 5352   region_13: 2648
##  10001  :    1   Finance          : 2536   region_26: 2260
##  (Other):54802   (Other)          : 4456   (Other)  :23478
##             education        gender               recruitment_channel
##                  : 2409   f     :16312   other              :30446
##  Bachelor's      :36669   gender:    0   recruitment_channel:    0
##  Below Secondary :  805   m     :38496   referred           : 1142
##  education       :    0                  sourcing           :23220
##  Master's & above:14925
##
##
##  no_of_trainings      age                 previous_year_rating
##  1      :44378   30     : 3665                      : 4124
##  2      : 7987   31     : 3534   1                 : 6223
##  3      : 1776   32     : 3534   2                 : 4225
##  4      :  468   29     : 3405   3                 :18618
##  5      :  128   33     : 3210   4                 : 9877
##  6      :   44   28     : 3147   5                 :11741
##  (Other):   27   (Other):34313   previous_year_rating:    0
##  length_of_service       KPIs_met >80%        awards_won?    avg_training_score
##  3      : 7033     0            :35517   0          :53538   50     : 2716
##  4      : 6836     1            :19291   1          : 1270   49     : 2681
##  2      : 6684     KPIs_met >80%:    0   awards_won?:    0   48     : 2437
##  5      : 5832                                                51     : 2347
##  7      : 5551                                                60     : 2155
##  6      : 4734                                                59     : 2064
##  (Other):18138                                                (Other):40408
##       is_promoted
##  0          :50140
##  1          : 4668
##  is_promoted:    0
##
##
##
##
```

Visualize every variable with target variable

Education and Promotion

```
train %>%
    group_by(education, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=education, y=n, fill=is_promoted)) + geom_bar(stat='identity', position='dod
ge')
```

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

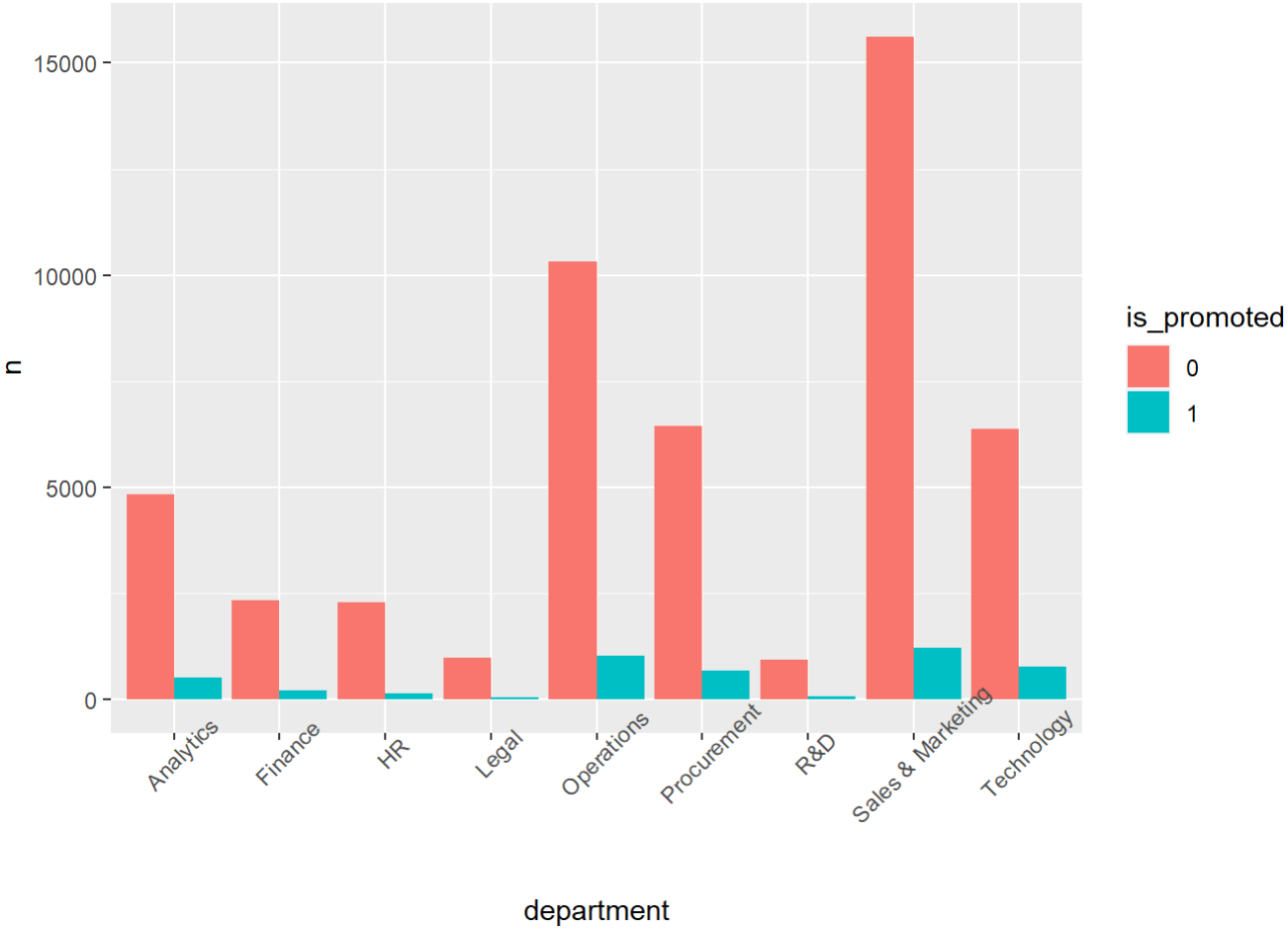


- In terms of education, promotion is focused on Bachelor's and above degree owners.

department and promotion

```
train %>%
    group_by(department, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=department, y=n, group=is_promoted, fill=is_promoted)) + geom_bar(stat='iden
tity', position='dodge') + theme(axis.text.x= element_text(angle=45))
```

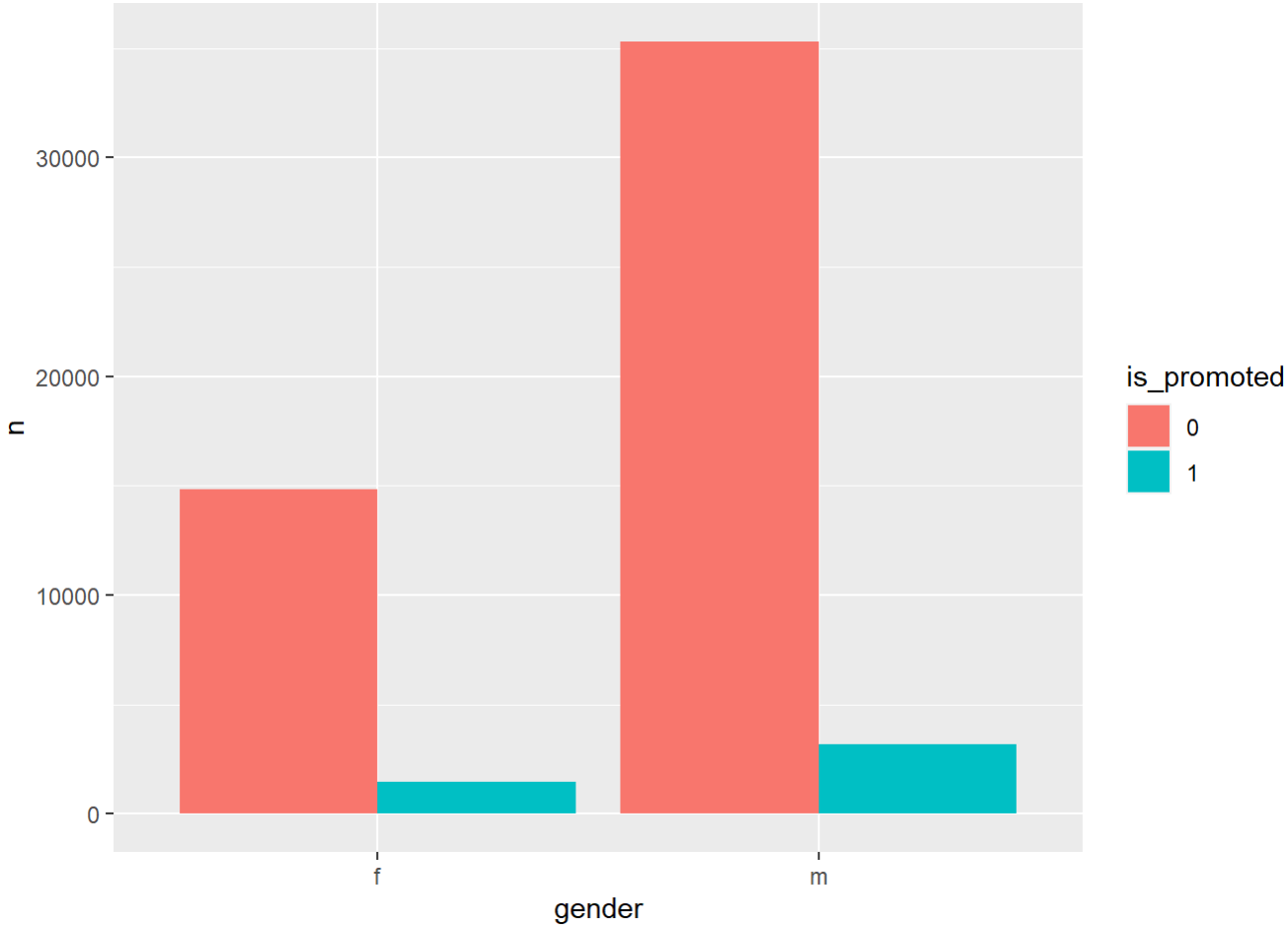
```
## `summarise()` has grouped output by 'department'. You can override using the
## `.groups` argument.
```



Gender and promotion

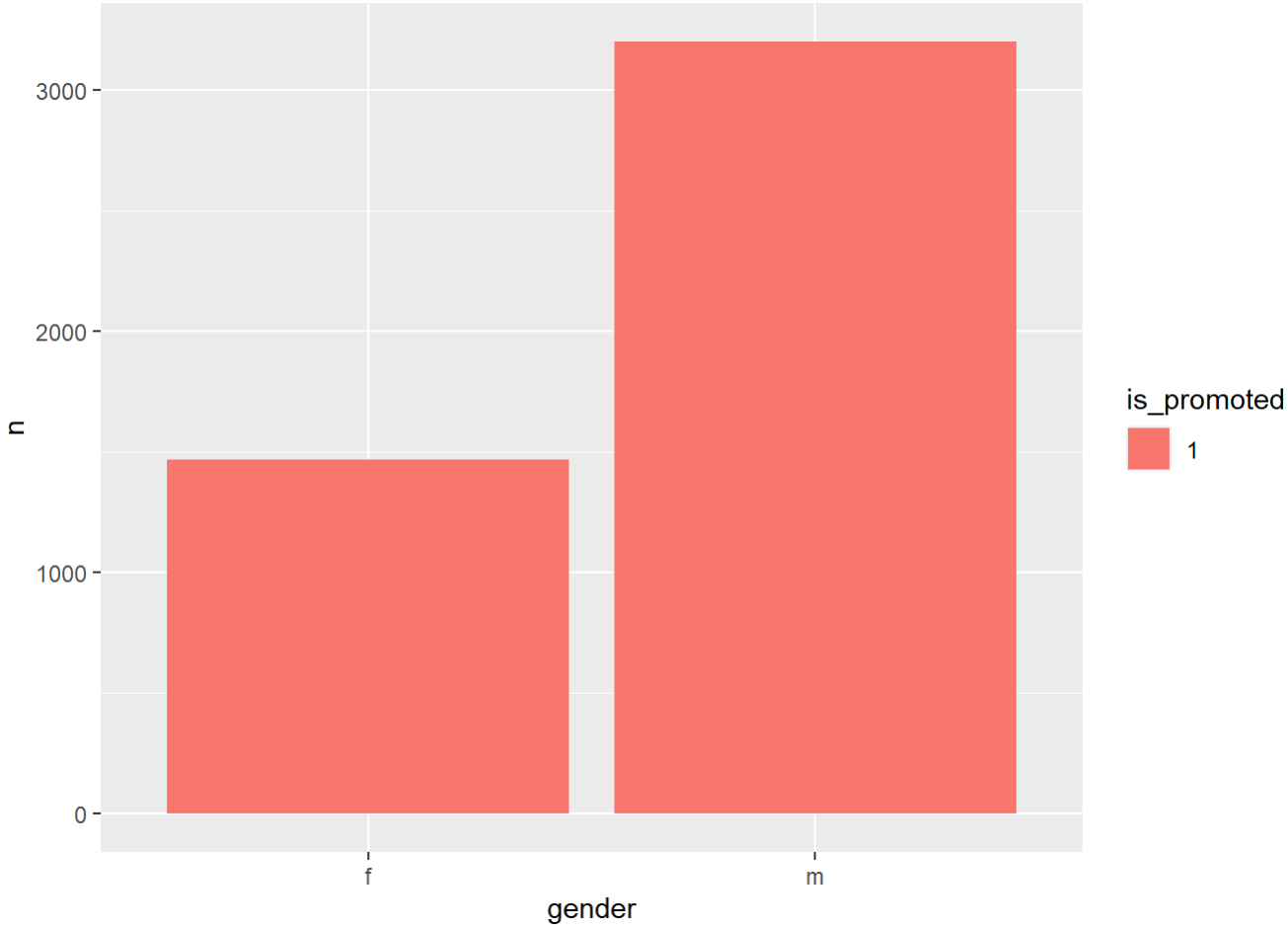
```
train %>%
    group_by(gender, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=gender, y=n, fill=is_promoted)) + geom_bar(stat='identity', position='dodg
e')
```

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




```
train %>%
    group_by(gender, is_promoted) %>%
    summarise(n=n()) %>%
    filter(is_promoted==1) %>%
    ggplot(aes(x=gender, y=n, fill=is_promoted)) + geom_bar(stat='identity', position='dodg
e')
```

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

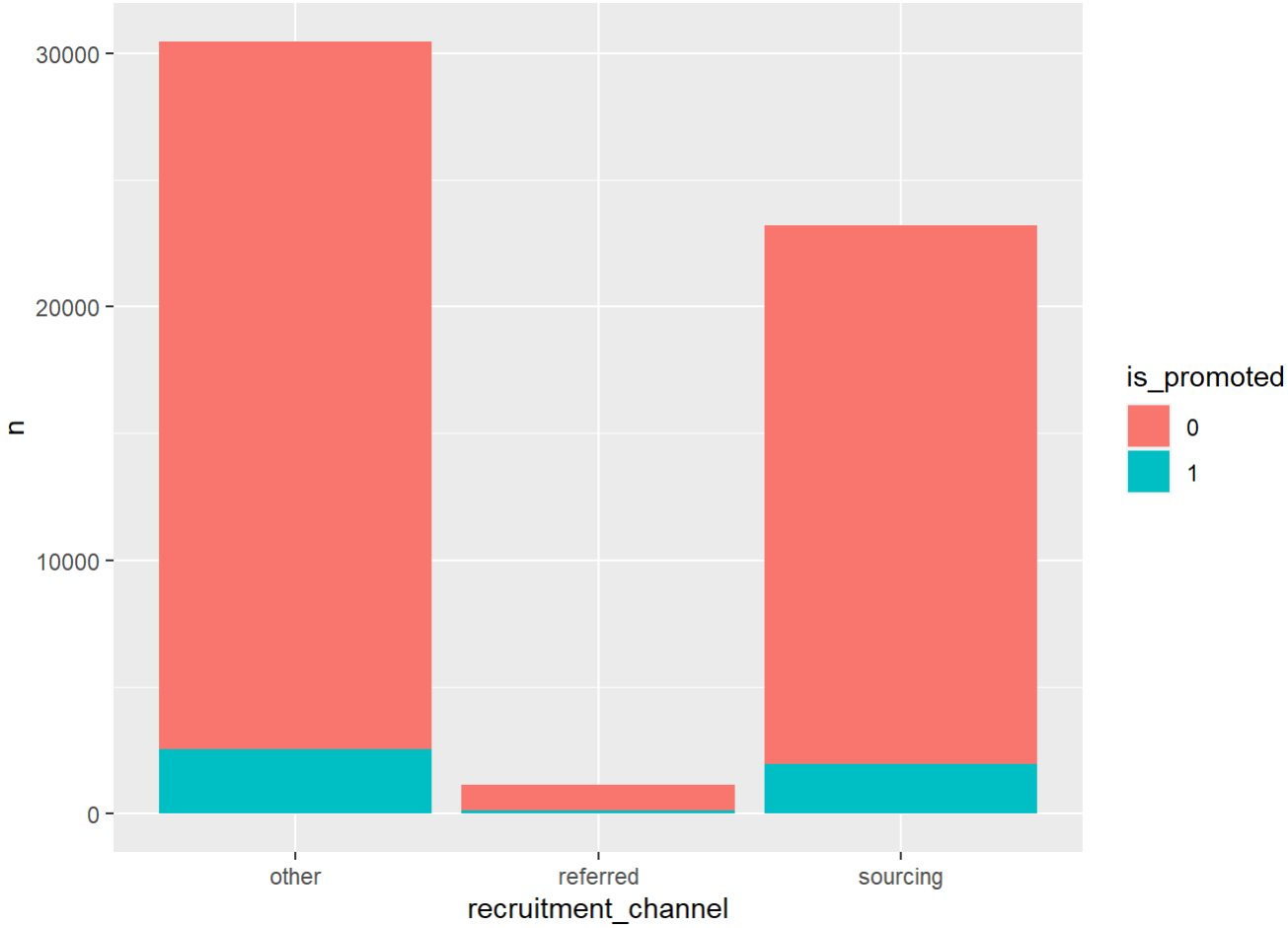


- The ratio of gender among employees is nearly identical to the ratio of gender among promotions.

recruitment_channel and promotion

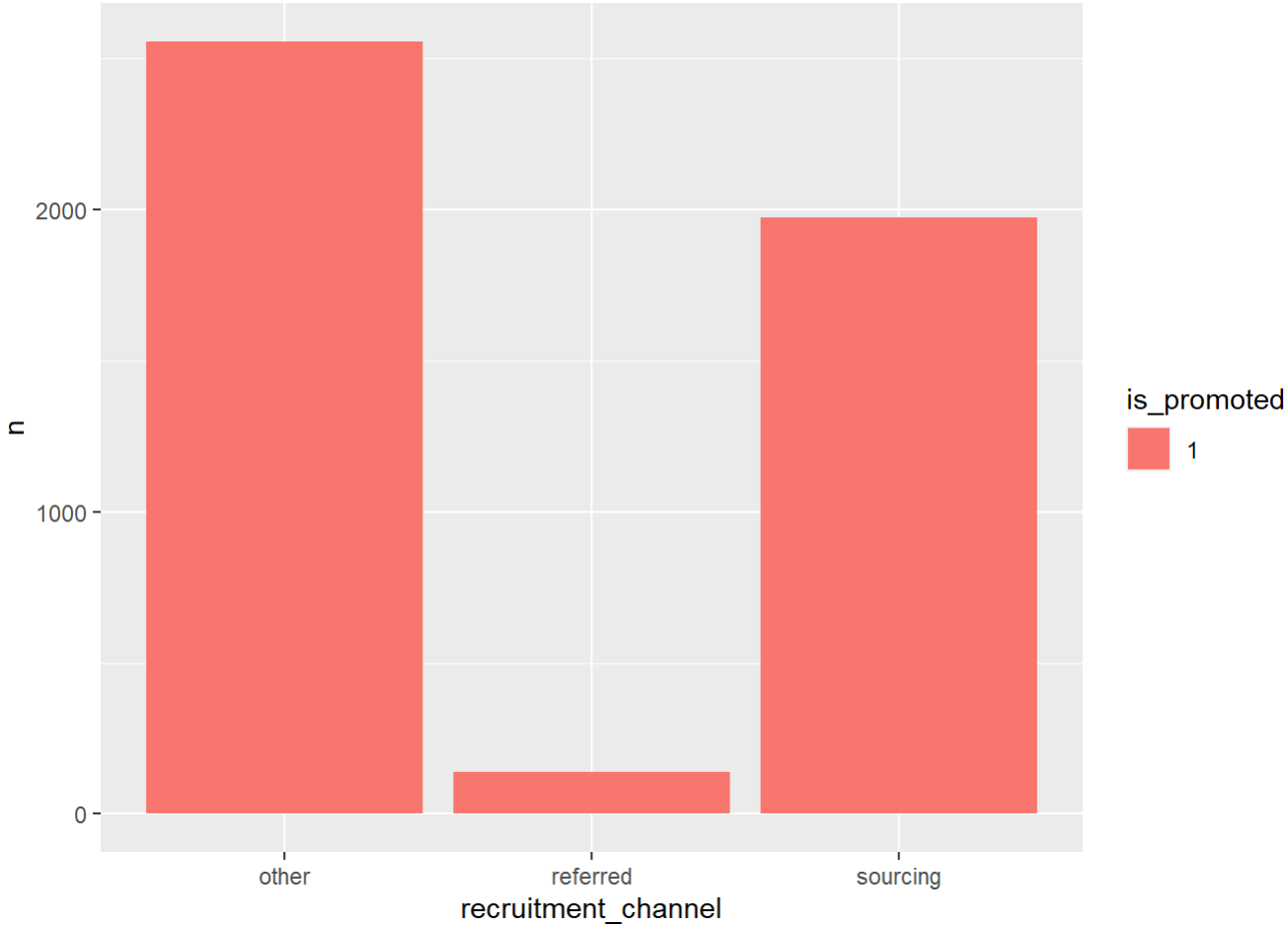
```
train %>%
    group_by(recruitment_channel, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=recruitment_channel, y=n, fill=is_promoted)) + geom_bar(stat='identity')
```

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




```
train %>%
    group_by(recruitment_channel, is_promoted) %>%
    summarise(n=n()) %>%
    filter(is_promoted==1) %>%
    ggplot(aes(x=recruitment_channel, y=n, fill=is_promoted)) + geom_bar(stat='identity')
```

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

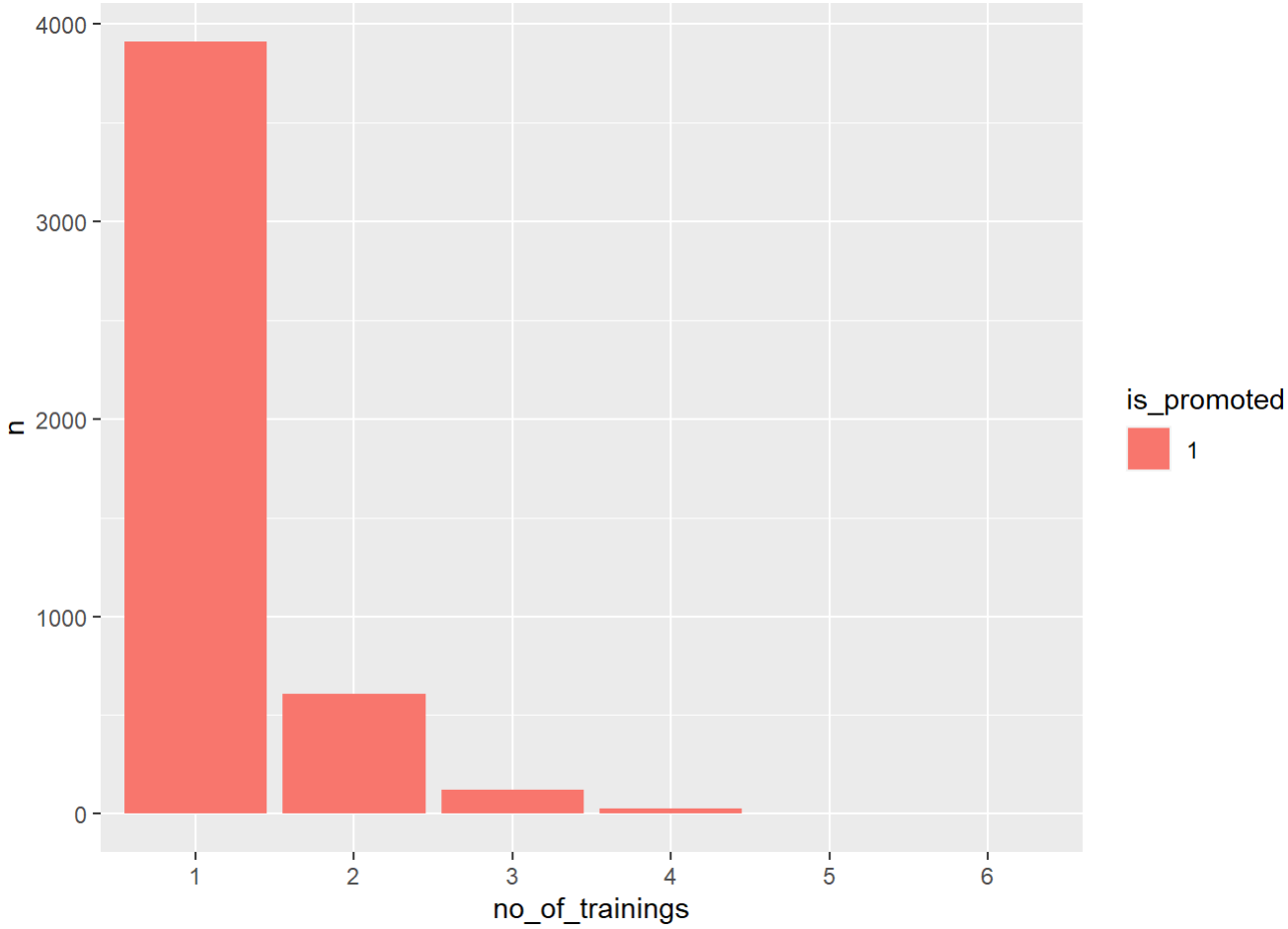


- Recruitment channel is not decisive factor to be promoted.

no_of_trainings and promotion

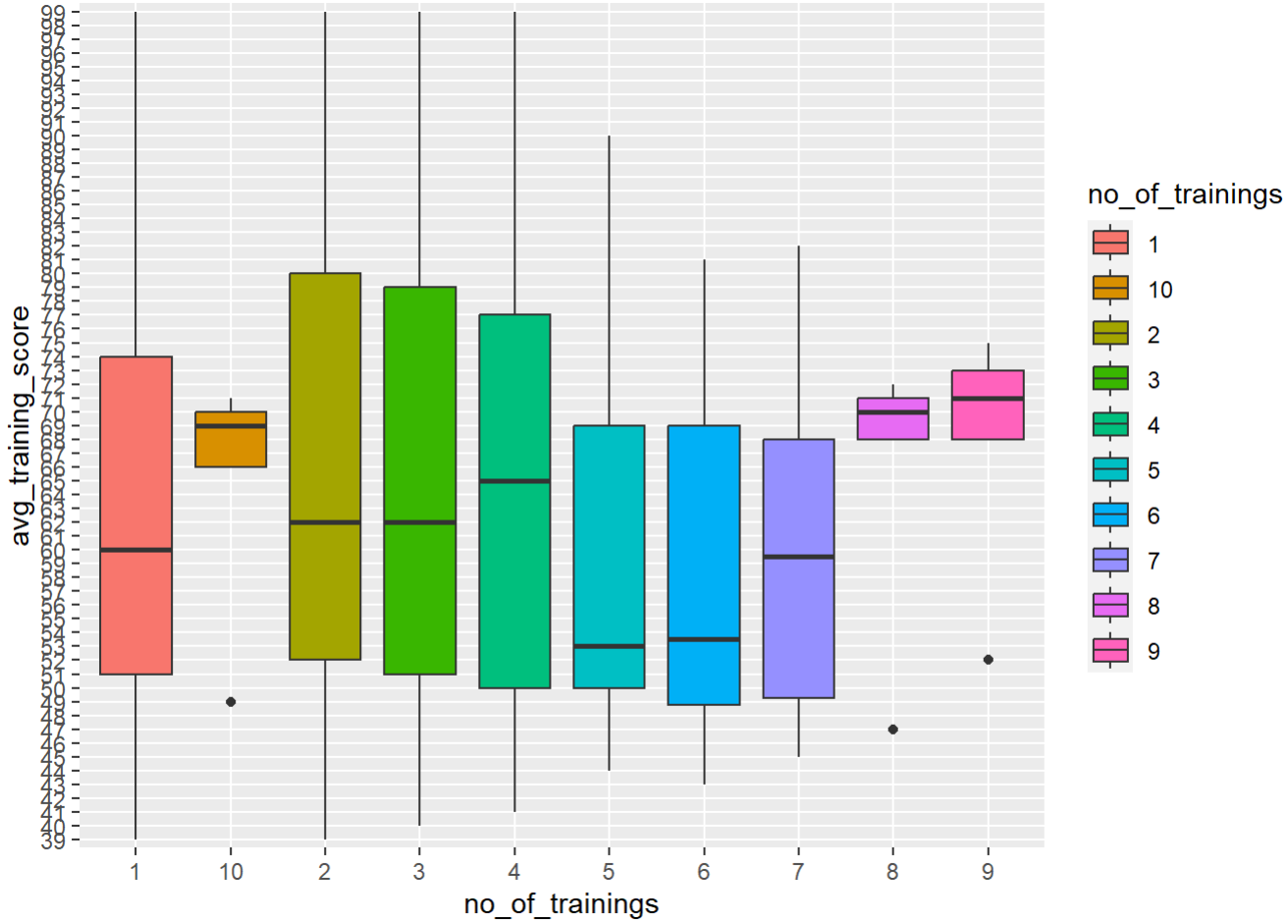
```
train %>%
    group_by(no_of_trainings, is_promoted) %>%
    summarise(n=n()) %>%
    filter(is_promoted==1) %>%
    ggplot(aes(x=no_of_trainings, y=n, fill=is_promoted)) + geom_bar(stat='identity',position
='dodge')
```

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




```
# no_of_trainings and avg_training_score
train %>%
    summarise(no_of_trainings=no_of_trainings, avg_training_score=avg_training_score) %>%
    ggplot(aes(x=no_of_trainings, y=avg_training_score, group=no_of_trainings, fill=no_of_tra
inings)) + geom_boxplot()
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



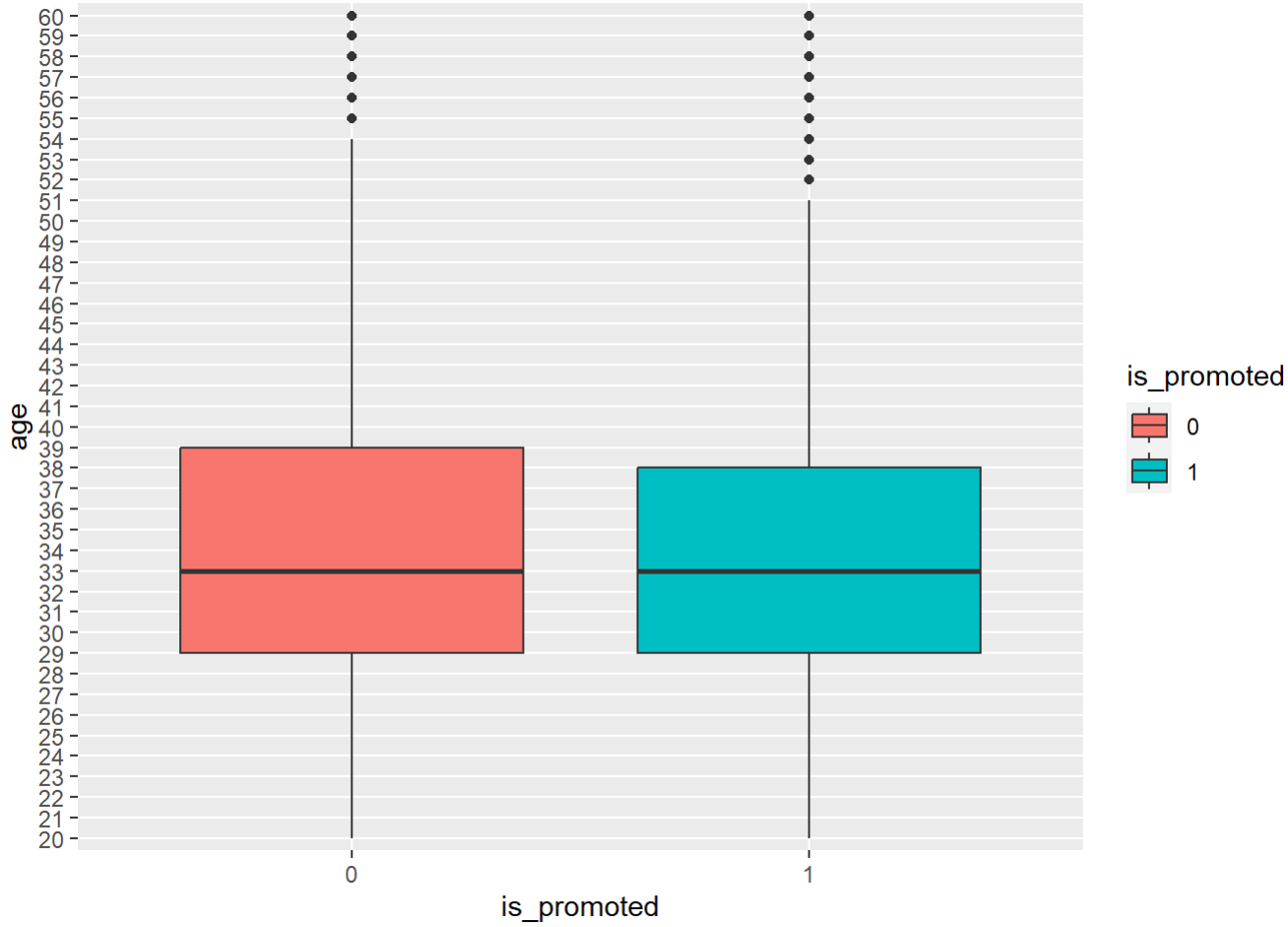
- The higher you gets score in training session, The more chances you have in promotion opportunity.

Age and promotion

```
train %>%
    group_by(is_promoted) %>%
    summarise(age=age, n=n()) %>%
    arrange(age) %>%
    ggplot(aes(x=is_promoted, y=age, group=is_promoted, fill=is_promoted)) + geom_boxplot()
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

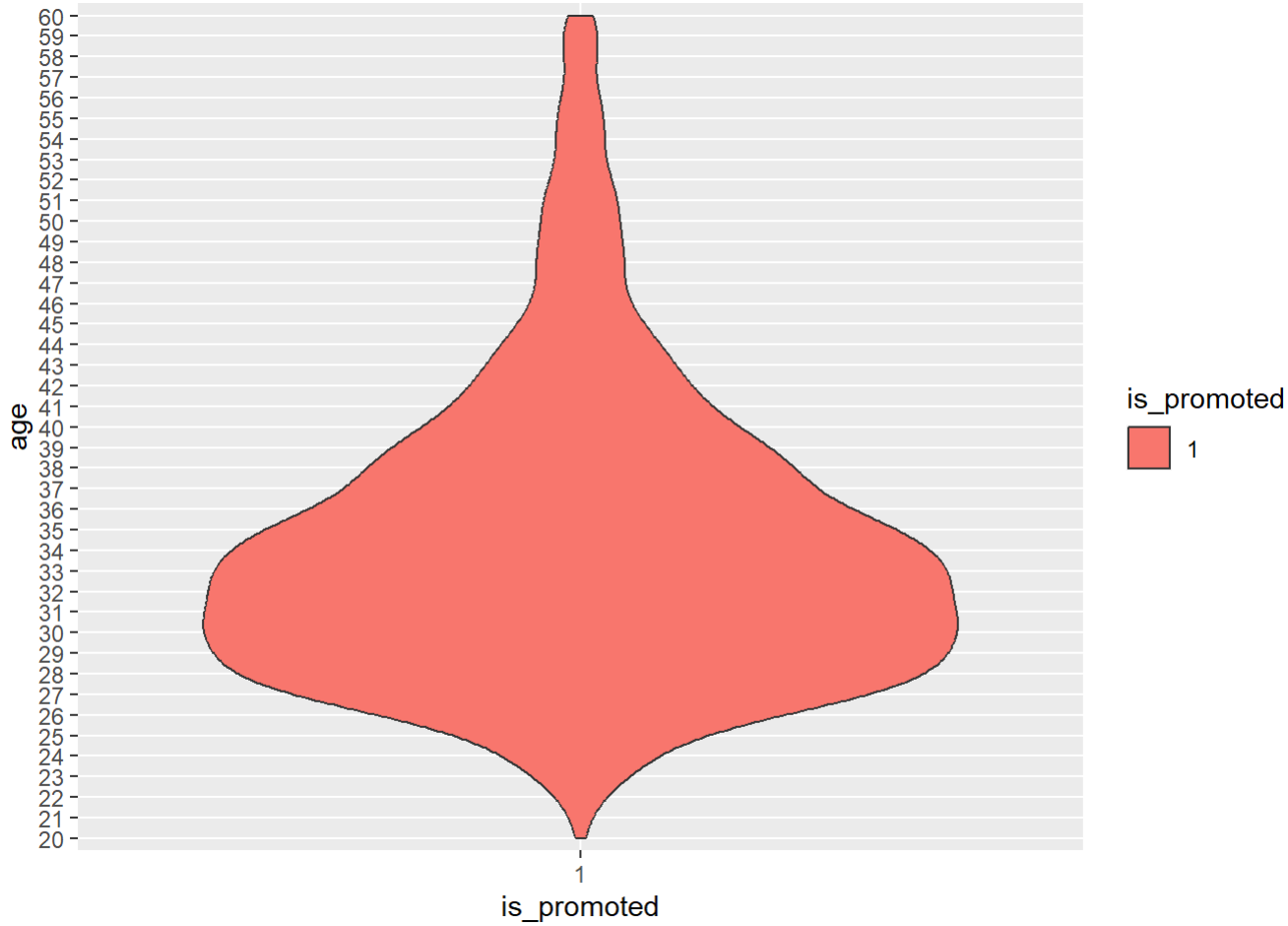
```
## `summarise()` has grouped output by 'is_promoted'. You can override using the
## `.groups` argument.
```




```
train %>%
    group_by(is_promoted) %>%
    summarise(age=age, n=n()) %>%
    arrange(age) %>%
    filter(is_promoted==1) %>%
    ggplot(aes(x=is_promoted, y=age, group=is_promoted, fill=is_promoted)) + geom_violin()
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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

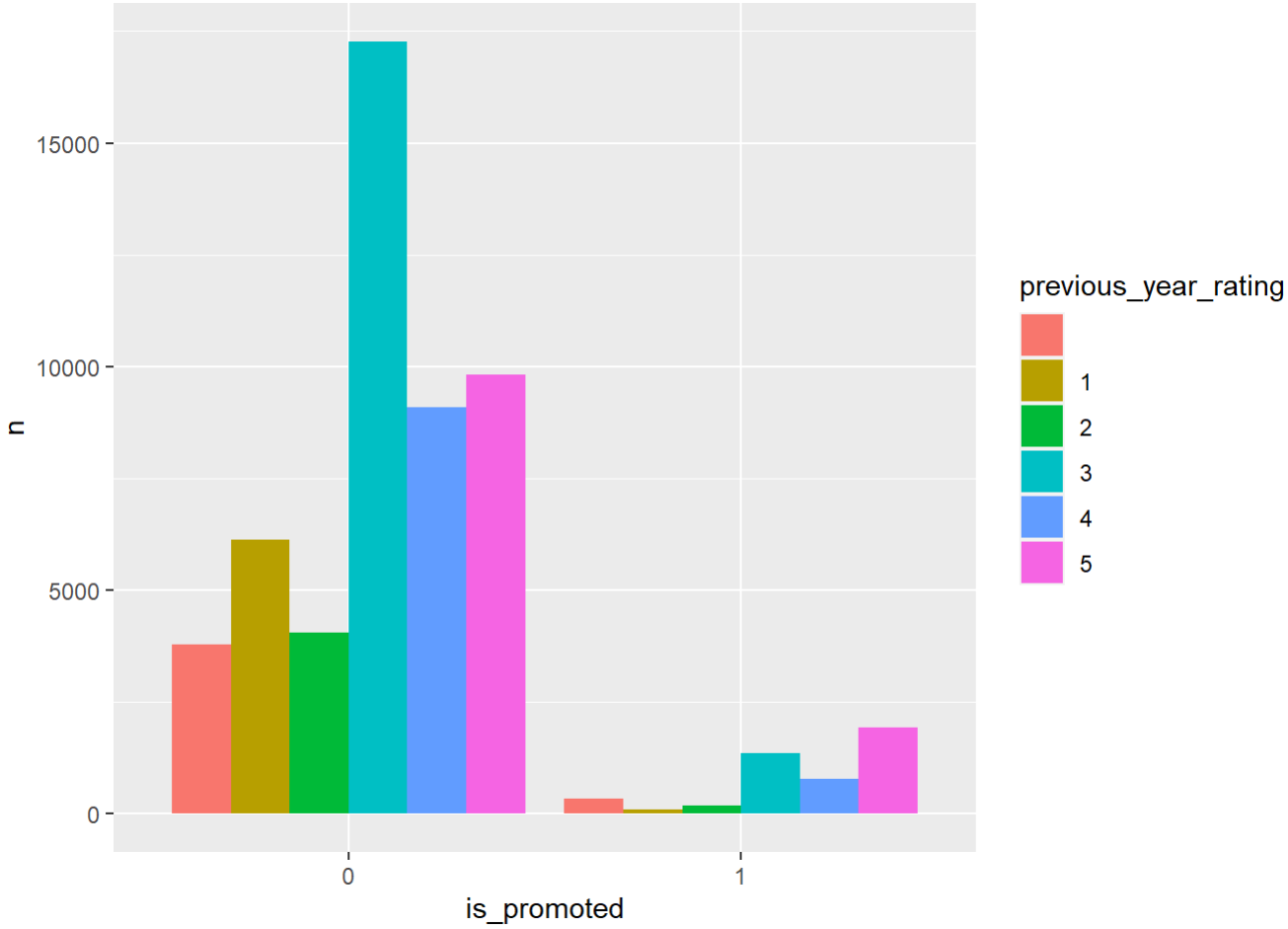


- Promotion is focused on early 30s, and the chance is drastically decreased for over mid 40 years employees.

previous year rating and promotion

```
train %>%
    group_by(previous_year_rating, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=is_promoted, y=n, group=previous_year_rating, fill=previous_year_rating)) +
geom_bar(stat='identity', position='dodge')
```

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



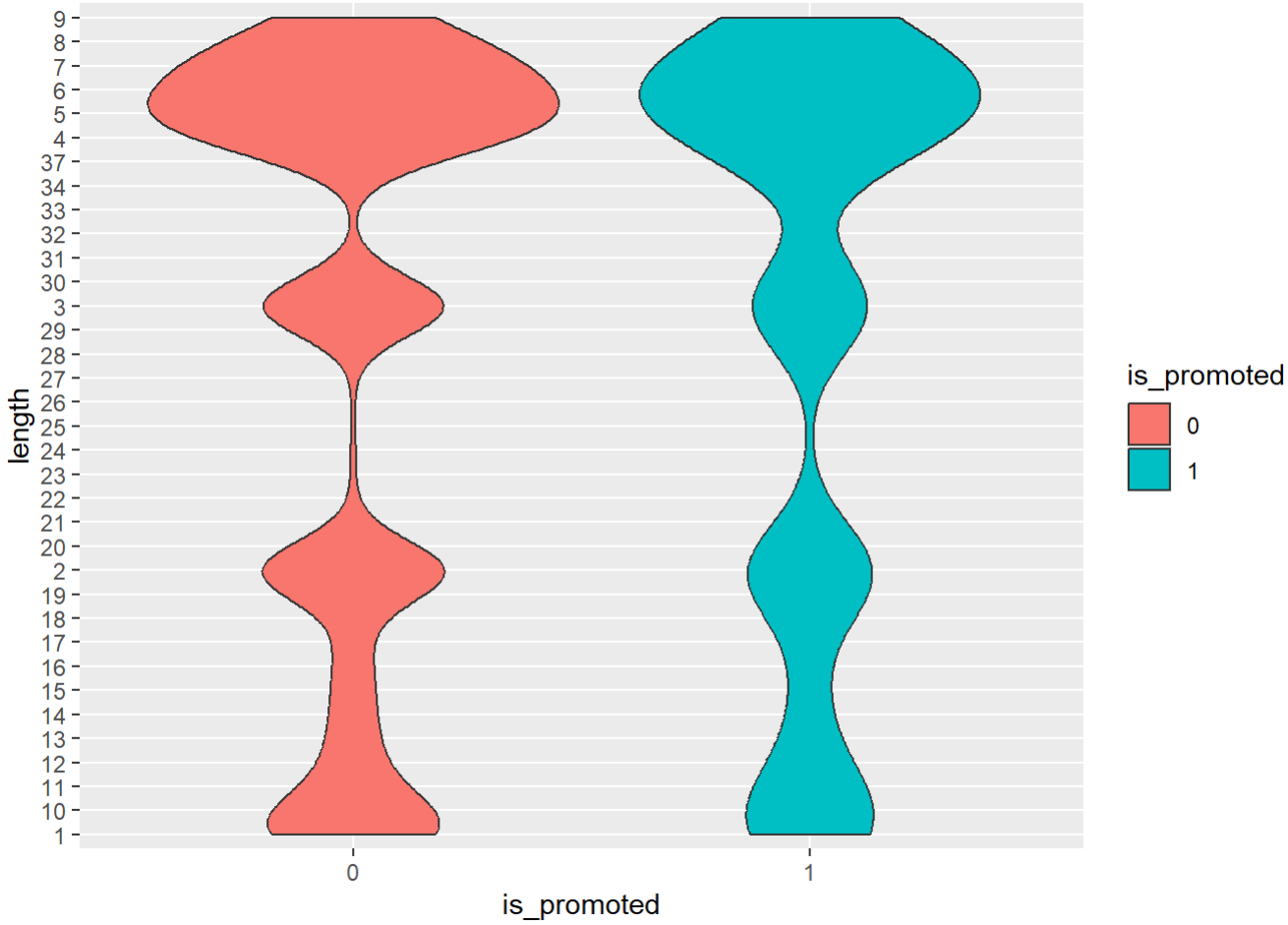
- Employees who got rated 1, 2 are merely promoted, but, in terms of ratio, employees who got rated 5 show highest promotion rate.

Length of service and promotion

```
train %>%
    group_by(is_promoted) %>%
    summarise(length=length_of_service) %>%
    #filter(is_promoted==1) %>%
    ggplot(aes(x=is_promoted, y=length, group=is_promoted, fill=is_promoted)) + geom_violin()
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



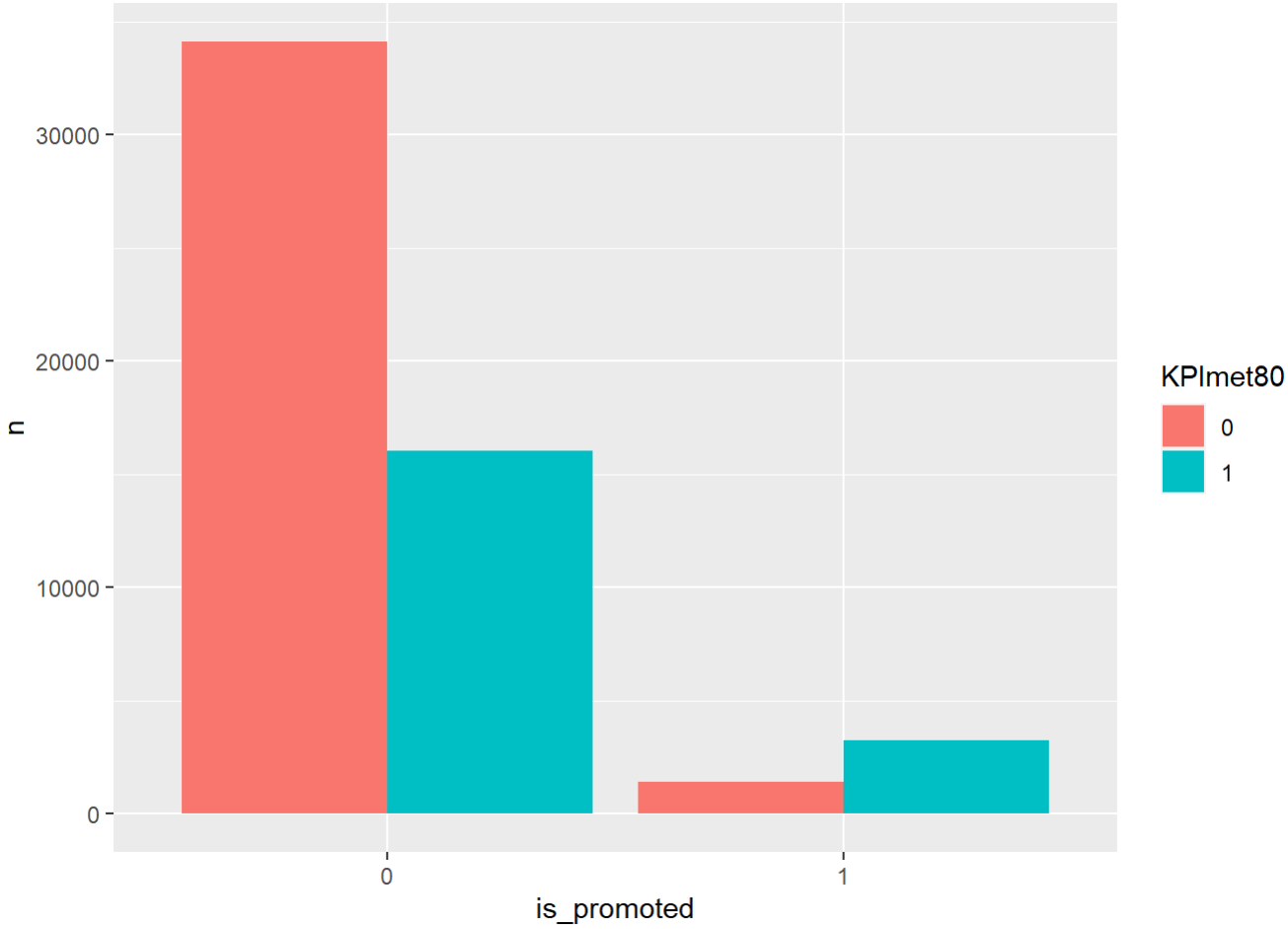
- Most promotion is focused on employees under 10 year length of services.

KPIs met >80% and promotion

```
colnames(train)[11] <- 'KPImet80'

train %>%
    group_by(KPImet80, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=is_promoted, y=n, group=KPImet80, fill=KPImet80)) + geom_bar(stat='identit
y', position='dodge')
```

```
## `summarise()` has grouped output by 'KPImet80'. You can override using the
## `.groups` argument.
```



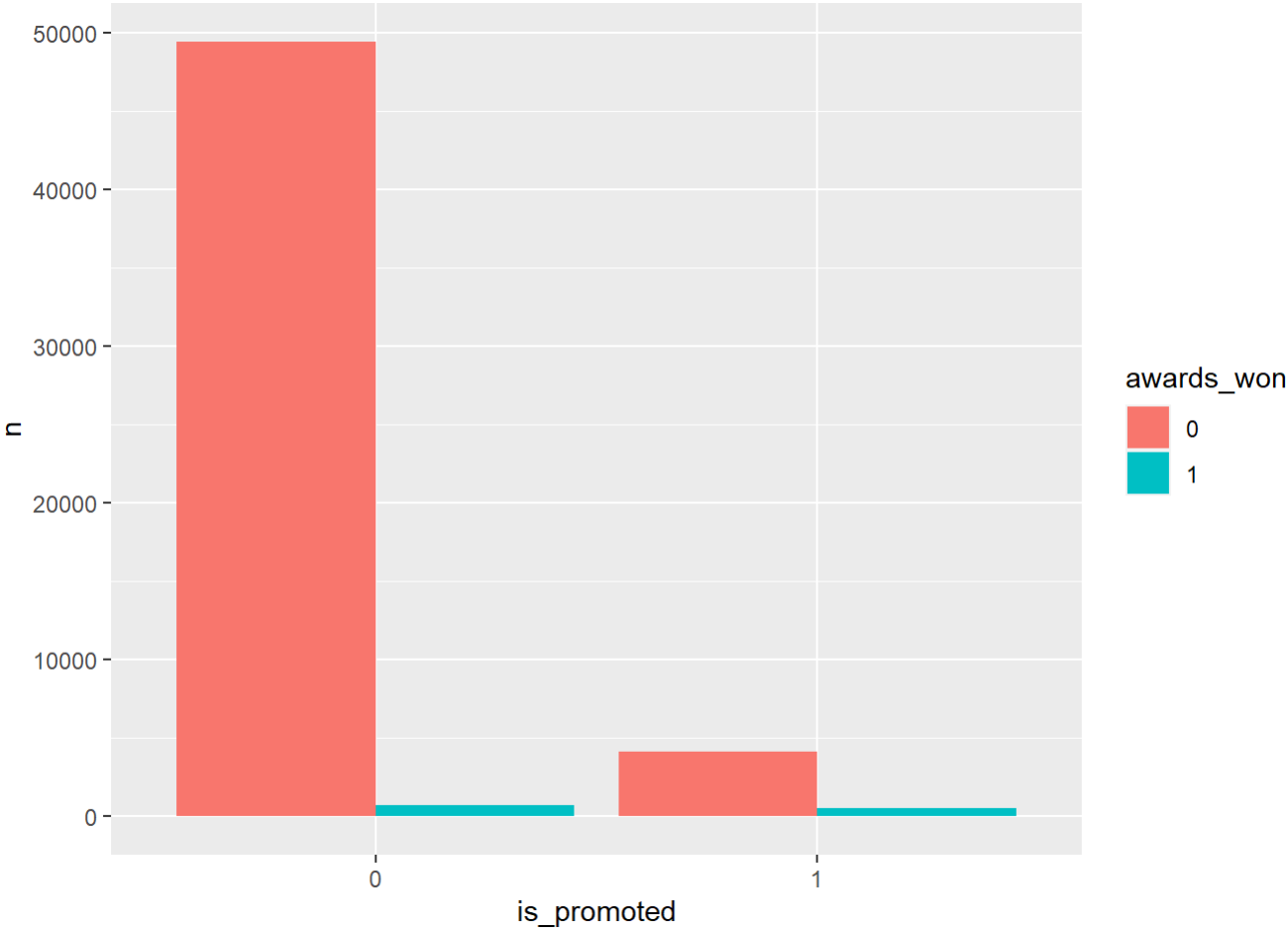
- When you met KPI requirements over 80 %, you make you chance of promotion twice.

awards_won and promotion

```
colnames(train)[12] <- 'awards_won'

train %>%
    group_by(awards_won, is_promoted) %>%
    summarise(n=n()) %>%
    ggplot(aes(x=is_promoted, y=n, group=awards_won, fill=awards_won)) + geom_bar(stat='ident
ity', position='dodge')
```

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



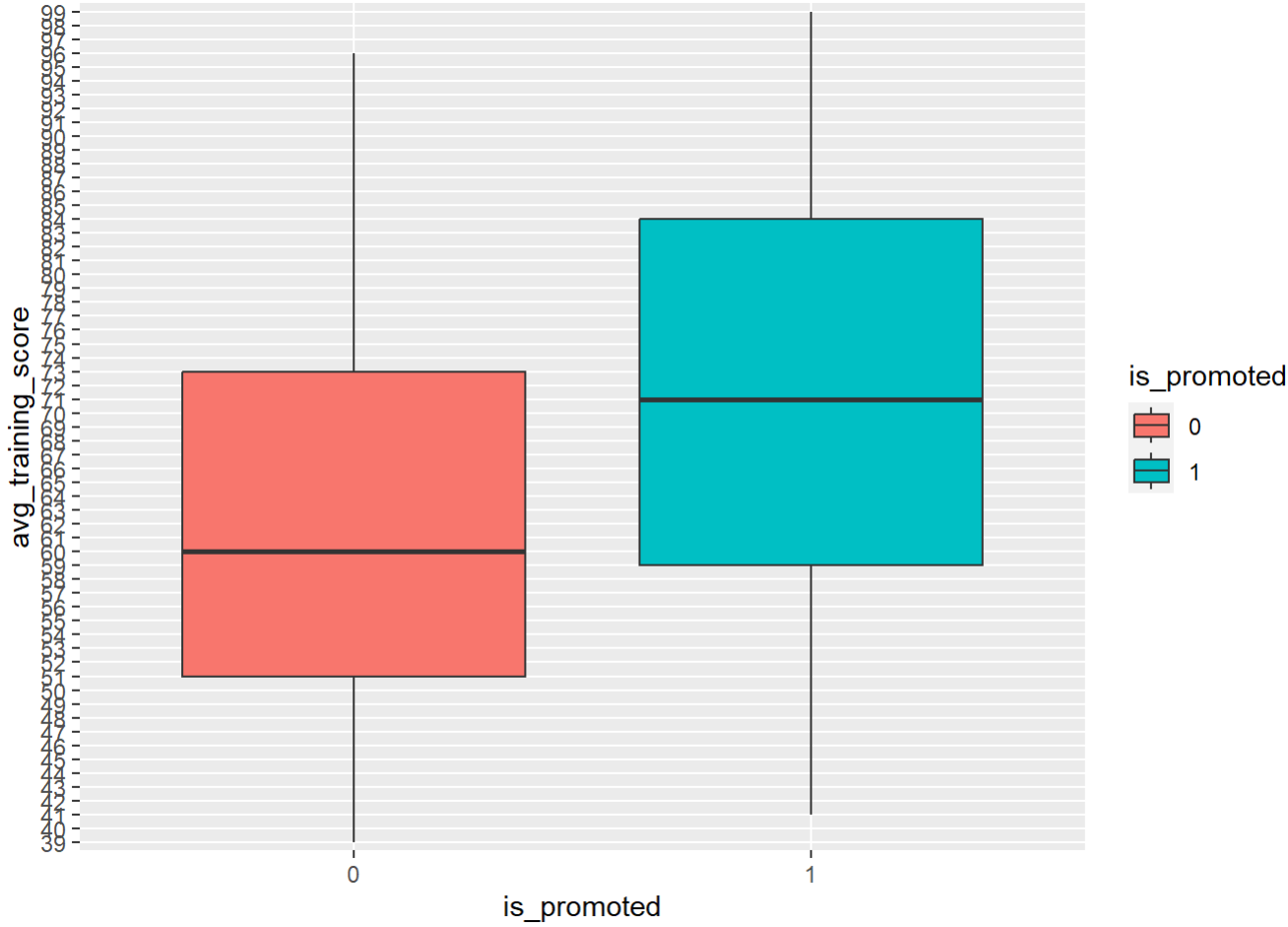
- Awards record is not a decisive factor for employees to be promoted.

avg_training_score and promotion

```
train %>%
    group_by(is_promoted) %>%
    summarise(avg_training_score=avg_training_score) %>%
    ggplot(aes(x=is_promoted, y=avg_training_score, group=is_promoted, fill=is_promoted)) + g
eom_boxplot()
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



- To be promoted, achieving at least over 70 training scores is recommended.

Step 4 : Summary & Recommendation

1. Among 8.5% of promotion chances, for employees want to be promoted, as an analyst, I recommend

- To get Bachelor's degree or above.
- To record over 70 % of training score on your the very first training session.
- To notice sales & marketing, operations, procurement, technology, and analytics department have top 5 promotion spots.
- To remind that most promotion chances are focused on 30 years old employees, and mid-40 is age limit.
- To achieve at least 3 from previous year rating, and over 4 ratings has more chances.
- To get over 80% of rating of KPI requirements.

2. For employers want to promote competent employees in efficient system

- Award is a irrelevant factor for promotion, so I recommend that setting a new and efficient system for awards and promotion.
- Salary is not on data. it caused a severe problem to analyze competence and compensations including promotion and salary.
- In recruitment channel, 'Other' has too many cases. I recommend to classify it in detail.
- The less employees take training session, the more employees got promoted. I think training session is changed into testing session. I recommend to training session must be focused on train employees, not assess them.