Exercise 2

Konstantin Volodin

2023-03-25

CODE

```
## Rows: 32906 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): application_number
## dbl  (2): ego_examiner_id, alter_examiner_id
## date (1): advice_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Proceeding with last name predictions...
##
## i All local files already up-to-date!
##
## i All local files already up-to-date!
##
## 701 (18.4%) individuals' last names were not matched.
```

Summary Statistics

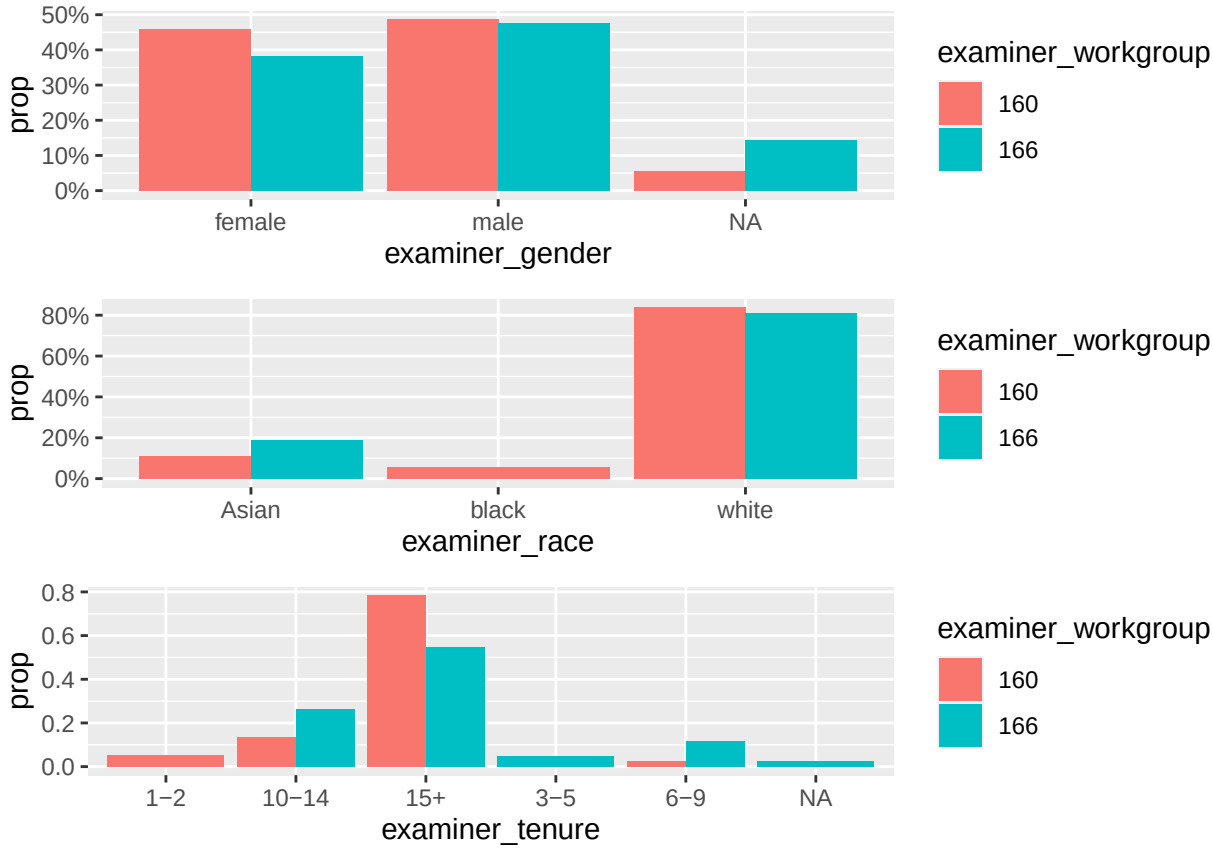
Workgroups 166 and 160 are evaluated. Workgroup 166 has less women, it has more asian, and it has significantly less employees with 15+ years of tenure.

```
## # A tibble: 2 x 4
## # Groups:   examiner_workgroup [2]
##   examiner_workgroup female  male  `NA`
##   <chr>               <dbl> <dbl> <dbl>
## 1 160                  45.9  48.6  5.41
## 2 166                  38.1  47.6 14.3
```

```
## # A tibble: 2 x 4
## # Groups:   examiner_workgroup [2]
##   examiner_workgroup Asian black white
##   <chr>              <dbl> <dbl> <dbl>
## 1 160                 10.8  5.41  83.8
## 2 166                 19.0 NA     81.0
```

```
## # A tibble: 2 x 7
## # Groups:   examiner_workgroup [2]
```

```
##   examiner_workgroup ‘1-2‘ ‘10-14‘ ‘15+‘ ‘6-9‘ ‘3-5‘  ‘NA‘
##   <chr>              <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 160                 5.41    13.5  78.4  2.70 NA    NA
## 2 166                NA       26.2  54.8 11.9   4.76  2.38
```




```
library(here)
library(arrow)
library(gender)
library(wru)
library(lubridate)

library(tidyverse)
library(igraph)
library(tidygraph)
library(ggraph)
library(gridExtra)


### LOAD DATA
applications <- read_parquet(here('assignments','assignment_3',"app_data_sample.parquet"))
edges <- read_csv(here('assignments','assignment_3',"edges_sample.csv"))


### GENDER
examiner_names <- applications %>% distinct(examiner_name_first)
examiner_names_gender <- examiner_names %>%
```

```
  do(results = gender(.$examiner_name_first, method = "ssa")) %>%
  unnest(cols = c(results), keep_empty = TRUE) %>%
  select( examiner_name_first = name, gender)
applications <- applications %>% left_join(examiner_names_gender, by = "examiner_name_first")


### RACE
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct(surname)
examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>%
  as_tibble() %>%
  mutate(max_race_p = pmax(pred.asi, pred.bla, pred.his, pred.oth, pred.whi)) %>%
  mutate(race = case_when(
    max_race_p == pred.asi ~ "Asian", max_race_p == pred.bla ~ "black",
    max_race_p == pred.his ~ "Hispanic", max_race_p == pred.oth ~ "other",
    max_race_p == pred.whi ~ "white", TRUE ~ NA_character_
  )) %>%
  select(surname,race)
applications <- applications %>% left_join(examiner_race, by = c("examiner_name_last" = "surname"))


### TENURE
examiner_dates <- applications %>% select(examiner_id, filing_date, appl_status_date)
examiner_dates <- examiner_dates %>%
  mutate(start_date = ymd(filing_date), end_date = as_date(dmy_hms(appl_status_date))) %>%
  group_by(examiner_id) %>%
  summarise(
    earliest_date = min(start_date, na.rm = TRUE),
    latest_date = max(end_date, na.rm = TRUE),
    tenure_days = interval(earliest_date, latest_date) %/% days(1)
  ) %>%
  filter(year(latest_date)<2018) %>%
  mutate(tenure_years = tenure_days / 365) %>%
    mutate(tenure = case_when(
      tenure_years <= 1 ~ '<1',
      tenure_years <= 2 ~ '1-2',
      tenure_years <= 5 ~ '3-5',
      tenure_years <= 9 ~ '6-9',
      tenure_years <= 14 ~ '10-14',
      tenure_years <= 100 ~ '15+',
      TRUE ~ NA_character_
    ))
applications <- applications %>% left_join(examiner_dates, by = "examiner_id")


### WORKGROUPS
applications <- applications %>% mutate(examiner_workgroup = str_sub(examiner_art_unit, 1, -2))


### CLEAN UP
rm(examiner_dates, examiner_names, examiner_names_gender, examiner_race, examiner_surnames)
```

```r
### EXAMINER DATA
examiner_data <- applications %>%
  distinct(examiner_id, examiner_gender = gender, examiner_race = race, examiner_tenure = tenure)


### WORKGROUPS
examiner_subset <- applications %>% filter(examiner_workgroup %in% c(166, 160)) %>%
  distinct(examiner_id, examiner_workgroup) %>%
  left_join(examiner_data, by='examiner_id')


### COMPARE WORKGROUPS (STATISTICS)
t_gend <- examiner_subset %>% count(examiner_workgroup, examiner_gender) %>%
  group_by(examiner_workgroup) %>% mutate(freq = n / sum(n) * 100) %>%
  select(examiner_workgroup, examiner_gender, freq) %>% pivot_wider(names_from = examiner_gender, value
t_race <- examiner_subset %>% count(examiner_workgroup, examiner_race) %>%
  group_by(examiner_workgroup) %>% mutate(freq = n / sum(n) * 100) %>%
  select(examiner_workgroup, examiner_race, freq) %>% pivot_wider(names_from = examiner_race, values_fr
t_tenure <- examiner_subset %>% count(examiner_workgroup, examiner_tenure) %>%
  group_by(examiner_workgroup) %>% mutate(freq = n / sum(n) * 100) %>%
  select(examiner_workgroup, examiner_tenure, freq) %>% pivot_wider(names_from = examiner_tenure, value


### COMPARE WORKGROUPS (PLOTS)
p_gend <- ggplot(examiner_subset, aes(x=examiner_gender, y=..prop.., fill=examiner_workgroup, group=exa
  geom_bar(aes(), stat='count', position='dodge') +
  scale_y_continuous(labels = scales::percent_format())
p_race <- ggplot(examiner_subset, aes(x=examiner_race, y=..prop.., fill=examiner_workgroup, group=exami
  geom_bar(aes(), stat='count', position='dodge') +
  scale_y_continuous(labels = scales::percent_format())
p_tenure <- ggplot(examiner_subset, aes(x=examiner_tenure, y=..prop.., fill=examiner_workgroup, group=e
  geom_bar(aes(), stat='count', position='dodge')


### CREATE NETWORK
edge_subset <- edges %>%
  filter(ego_examiner_id %in% examiner_subset$examiner_id |
           alter_examiner_id %in% examiner_subset$examiner_id) %>%
  drop_na() %>%
  select(from = ego_examiner_id, to = alter_examiner_id)
# nodes <-
node_subset <- edge_subset %>%
  pivot_longer(cols=c('from','to')) %>%
  distinct(examiner_id = value) %>%
  # left_join(examiner_data, on='examiner_id') %>%
  # distinct(examiner_id, examiner_gender, examiner_race, examiner_tenure) %>%
  rename(name = examiner_id)
network <- tbl_graph(edges = edge_subset)


### ESTIMATE METRICS
# network <- network %>%
#   mutate(degree = centrality_degree(),
```

```
#          closeness = centrality_closeness_harmonic(),
#          betweenness = centrality_betweenness()) %>%
#   mutate(avg = (degree + closeness + betweenness)/3) %>%
#   mutate(label = paste0(name, '\n',
#                         'Degree: ',round(degree,2), '\n',
#                         'Closeness: ',round(closeness,2), '\n',
#                         'Betweenness: ',round(betweenness,2), '\n',
#                         'Avg: ',round(avg,2)))
# node_data <- network %>% data.frame() %>% tibble()
#

### PLOT NETWORK
```