exercise4

jingwen he

2024-04-09

```
 options(repos = c(CRAN = "https://cran.rstudio.com/"))

install.packages("tinytex")
```

```
## Installing package into 'C:/Users/17286/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'tinytex' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17286\AppData\Local\Temp\RtmpE5DJ1a\downloaded_packages
```

```
tinytex::install_tinytex(force = TRUE)
```

```
## tlmgr install tlgpg

## tlmgr update --self

## tlmgr install tlgpg

## tlmgr --repository http://www.preining.info/tlgpg/ install tlgpg

## tlmgr option repository "https://ctan.mirror.globo.tech/systems/texlive/tlnet"

## tlmgr update --list
```

```
tinytex:::is_tinytex()
```

```
## [1] TRUE
```

```
install.packages("arrow")
```

```
## Installing package into 'C:/Users/17286/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'arrow' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17286\AppData\Local\Temp\RtmpE5DJ1a\downloaded_packages
```

```r
library(arrow)
```

```
##
## Attaching package: 'arrow'
```

```
## The following object is masked from 'package:utils':
##
##     timestamp
```

```r
file_path <- "C:/Users/17286/Downloads/app_data_sample.parquet"
data <- read_parquet(file_path)
head(data)
```

```
##   application_number filing_date examiner_name_last examiner_name_first
## 1           08284457  2000-01-26             HOWARD          JACQUELINE
## 2           08413193  2000-10-11           YILDIRIM               BEKIR
## 3           08531853  2000-05-17           HAMILTON             CYNTHIA
## 4           08637752  2001-07-20             MOSHER                MARY
## 5           08682726  2000-04-10               BARR             MICHAEL
## 6           08687412  2000-04-28               GRAY               LINDA
##   examiner_name_middle examiner_id examiner_art_unit uspc_class uspc_subclass
## 1                    V       96082              1764        508        273000
## 2                    L       87678              1764        208        179000
## 3                 <NA>       63213              1752        430        271100
## 4                 <NA>       73788              1648        530        388300
## 5                    E       77294              1762        427        430100
## 6                LAMEY       68606              1734        156        204000
##   patent_number patent_issue_date abandon_date disposal_type appl_status_code
## 1       6521570        2003-02-18         <NA>           ISS              150
## 2       6440298        2002-08-27         <NA>           ISS              250
## 3       5607816        1997-03-04         <NA>           ISS              250
## 4       6927281        2005-08-09         <NA>           ISS              250
## 5          <NA>              <NA>   2000-12-27           ABN              161
## 6       6267836        2001-07-31         <NA>           ISS              150
##     appl_status_date   tc
## 1 30jan2003 00:00:00 1700
## 2 27sep2010 00:00:00 1700
## 3 30mar2009 00:00:00 1700
## 4 07sep2009 00:00:00 1600
## 5 19apr2001 00:00:00 1700
## 6 16jul2001 00:00:00 1700
```

```r
str(data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2018477 obs. of  16 variables:
##  $ application_number  : chr  "08284457" "08413193" "08531853" "08637752" ...
##  $ filing_date         : Date, format: "2000-01-26" "2000-10-11" ...
##  $ examiner_name_last  : chr  "HOWARD" "YILDIRIM" "HAMILTON" "MOSHER" ...
##  $ examiner_name_first : chr  "JACQUELINE" "BEKIR" "CYNTHIA" "MARY" ...
##  $ examiner_name_middle: chr  "V" "L" NA NA ...
##  $ examiner_id         : num  96082 87678 63213 73788 77294 ...
```

```
##  $ examiner_art_unit   : num  1764 1764 1752 1648 1762 ...
##  $ uspc_class          : chr  "508" "208" "430" "530" ...
##  $ uspc_subclass       : chr  "273000" "179000" "271100" "388300" ...
##  $ patent_number       : chr  "6521570" "6440298" "5607816" "6927281" ...
##  $ patent_issue_date   : Date, format: "2003-02-18" "2002-08-27" ...
##  $ abandon_date        : Date, format: NA NA ...
##  $ disposal_type       : chr  "ISS" "ISS" "ISS" "ISS" ...
##  $ appl_status_code    : num  150 250 250 250 161 150 135 161 161 250 ...
##  $ appl_status_date    : chr  "30jan2003 00:00:00" "27sep2010 00:00:00" "30mar2009 00:00:00" "07sep20
##  $ tc                  : num  1700 1700 1700 1600 1700 1700 1600 1600 1600 1700 ...
```

```
#1
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:arrow':
##
##     duration

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# cahnge the type of the date
data$filing_date <- as.Date(data$filing_date)
data$patent_issue_date <- as.Date(data$patent_issue_date)
data$abandon_date <- as.Date(data$abandon_date)
# calculate the time
data$appproc_time <- ifelse(!is.na(data$patent_issue_date),
                            data$patent_issue_date - data$filing_date,
                            data$abandon_date - data$filing_date)
data$appproc_time_days <- as.numeric(data$appproc_time)
#2
library(igraph)
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:lubridate':
##
##     %--%, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```r
install.packages("tidyverse")
```

```
## Installing package into 'C:/Users/17286/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17286\AppData\Local\Temp\RtmpE5DJ1a\downloaded_packages
```

```r
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:igraph':
##
##     crossing
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(readr)
install.packages("wru")
```

```
## Installing package into 'C:/Users/17286/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'wru' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17286\AppData\Local\Temp\RtmpE5DJ1a\downloaded_packages
```

```
library(wru)
```

```
##
## Please cite as:
##
## Khanna K, Bertelsen B, Olivella S, Rosenman E, Rossell Hayes A, Imai K
## (2024). _wru: Who are You? Bayesian Prediction of Racial Category Using
## Surname, First Name, Middle Name, and Geolocation_. R package version
## 3.0.2, <https://CRAN.R-project.org/package=wru>.
##
## Note that wru 2.0.0 uses 2020 census data by default.
## Use the argument `year = "2010"`, to replicate analyses produced with earlier package versions.
```

```
library(lubridate)
library(tidyr)
library(gender)
file_path <- "C:/Users/17286/Downloads/edges_sample.csv" # Make sure the file path is correct and inclu
edges_sample <- read_csv(file_path)
```

```
## 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.
```

```
# Calculate the degree centrality
edges_sample = drop_na (edges_sample)
edges_sample = select (edges_sample,ego_examiner_id, alter_examiner_id)
library(igraph)
g <- graph_from_data_frame(edges_sample, directed = FALSE)
degree_centrality <- degree(g, mode="all")
centrality_df <- data.frame(examiner_id = V(g)$name, degree_centrality = degree_centrality)
data <- merge(data, centrality_df, by = "examiner_id")

#get the name, gender, and race
examiner_names <- data %>%
  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,
    proportion_female
  )
data <- data %>%
  left_join(examiner_names_gender, by = "examiner_name_first")
```

```r
rm(examiner_names)
rm(examiner_names_gender)
# get the surname and race
examiner_surnames <- data %>%
  select(surname = examiner_name_last) %>%
  distinct()
examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = TRUE) %>%
  as_tibble()
```

```
## Predicting race for 2020
## Proceeding with last name predictions...
## i All local files already up-to-date!
## 262 (15.5%) individuals' last names were not matched.
```

```r
examiner_race <- examiner_race %>%
  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_
  ))
data <- data %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))
rm(examiner_race)
rm(examiner_surnames)

# calculate the working time
examiner_dates <- data %>%
  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)))
examiner_dates <- examiner_dates %>%
  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)
data <- data %>%
  left_join(examiner_dates, by = "examiner_id")
rm(examiner_dates)

# use the lm
model_with_centrality <- lm(appproc_time_days ~ gender + race + degree_centrality + tenure_days, data =
summary(model_with_centrality)
```

```
##
## Call:
```

```
## lm(formula = appproc_time_days ~ gender + race + degree_centrality +
##     tenure_days, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7344.4  -426.4  -113.2   290.7  4959.9
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.513e+03  5.920e+00 255.608  < 2e-16 ***
## gendermale         1.250e+01  1.433e+00   8.722  < 2e-16 ***
## raceBlack         -3.014e+01  3.691e+00  -8.164 3.23e-16 ***
## raceHispanic       1.698e+01  4.704e+00   3.610 0.000307 ***
## raceOther          1.819e+02  1.941e+01   9.368  < 2e-16 ***
## raceWhite         -6.197e+01  1.498e+00 -41.359  < 2e-16 ***
## degree_centrality  3.893e-01  2.039e-02  19.092  < 2e-16 ***
## tenure_days       -4.626e-02  9.615e-04 -48.107  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 625.7 on 906623 degrees of freedom
##   (332087 observations deleted due to missingness)
## Multiple R-squared:  0.005717,   Adjusted R-squared:  0.00571
## F-statistic: 744.8 on 7 and 906623 DF,  p-value: < 2.2e-16
```

```r
# 3  add the interaction term
model <- lm(appproc_time_days ~ gender * degree_centrality + race , data = data)
summary(model)
```

```
##
## Call:
## lm(formula = appproc_time_days ~ gender * degree_centrality +
##     race, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10521.8   -427.5   -112.8    291.2   4954.3
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1231.95762    1.73209 711.257  < 2e-16 ***
## gendermale                     26.68976    1.64561  16.219  < 2e-16 ***
## degree_centrality               0.78394    0.04302  18.223  < 2e-16 ***
## raceBlack                     -32.94762    3.69608  -8.914  < 2e-16 ***
## raceHispanic                   21.65593    4.71331   4.595 4.34e-06 ***
## raceOther                     178.51117   19.45138   9.177  < 2e-16 ***
## raceWhite                     -65.08755    1.49131 -43.645  < 2e-16 ***
## gendermale:degree_centrality   -0.54532    0.04881 -11.172  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 626.9 on 914874 degrees of freedom
##   (323836 observations deleted due to missingness)
## Multiple R-squared:  0.003107,   Adjusted R-squared:  0.003099
```

```
## F-statistic: 407.3 on 7 and 914874 DF,  p-value: < 2.2e-16
```

```
## Based on the summary of the linear regression model with the interaction term gender * degree_centra

##Q4: Resource Allocation: The findings could inform resource and workload allocation at the USPTO. For
```

Based on the summary of the linear regression model with the interaction term gender * degree_centrality, we can determine the relationship between degree centrality and application processing time (app_proc_time_days) differs by gender. (The coefficient for gendermale is positive, indicating that male examiners have a higher app_proc_time_days compared to the baseline gender group)

##Q4: Resource Allocation: The findings could inform resource and workload allocation at the USPTO. For instance, if centrality (which might be a proxy for workload or network involvement) impacts processing times differently across genders, management might consider these dynamics to optimize performance. Gender Dynamics: There may be underlying gender dynamics that influence how work is processed and how connections are utilized in the workplace. The USPTO could further investigate the reasons behind these differences.