Untitled

```
# set path for R to find our data
data_path <- "C:/Users/Admin/Desktop/STATS/"
library(arrow) # to be able to load data in the .parquet format
```

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

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

```
# read application data
app_data_sample <- read_parquet(paste0(data_path,"app_data_sample.parquet"))
library(gender)
#install_genderdata_package() # only run this line the first time you use the package
# get a list of first names without repetitions
examiner_names <- app_data_sample %>%
  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
  )
# remove extra colums from the gender table
examiner_names_gender <- examiner_names_gender %>%
  select(examiner_name_first, gender)
# joining gender back to the dataset
app_data_sample <- app_data_sample %>%
  left_join(examiner_names_gender, by = "examiner_name_first")
# cleaning up
rm(examiner_names)
rm(examiner_names_gender)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4519925 241.4    7984011 426.4  4539915 242.5
## Vcells 49472030 377.5   95377472 727.7 79787788 608.8
```

```
# Examiners' race
library(wru)
examiner_surnames <- app_data_sample %>%
  select(surname = examiner_name_last) %>%
  distinct()
examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>%
  as_tibble()
```

```
## [1] "Proceeding with surname-only predictions..."

## Warning in merge_surnames(voter.file): Probabilities were imputed for 698
## surnames that could not be matched to Census list.
```

```
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_
  ))
examiner_race <- examiner_race %>%
  select(surname,race)
app_data_sample <- app_data_sample %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))
rm(examiner_race)
rm(examiner_surnames)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4934664 263.6    7984011 426.4  7984011 426.4
## Vcells 53271269 406.5   95377472 727.7 95170760 726.1
```

```
# Examiner's tenure
library(lubridate) # to work with dates
```

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

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

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

```
examiner_dates <- app_data_sample %>%
  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))
  filter(year(end_date)<2018) %>%
  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)
    )
app_data_sample <- app_data_sample %>%
  left_join(examiner_dates, by = "examiner_id")
rm(examiner_dates)
gc()
```

```
##            used  (Mb) gc trigger   (Mb)  max used (Mb)
## Ncells  4949513 264.4   14342938  766.0  14342938  766
## Vcells 65651540 500.9  165103470 1259.7 137489761 1049
```

Adding paygrade data

First, we load the paygrade file.

```
examiner_gs <- read_csv(paste0(data_path,"examiner_gs.csv"))
```

```
## Rows: 52109 Columns: 6
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): examiner_name, start_date, end_date
## dbl (3): examiner_grade, old_pid, new_pid
##
## 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.
```

```
examiner_ids <- read_csv(paste0(data_path,"examiner_ids.csv"))
```

```
## Rows: 19454 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): examiner_name
## dbl (3): old_pid, new_pid, patex_id
##
## 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.
```

```
examiner_gs <- examiner_gs %>%
  left_join(examiner_ids) %>%
  select(
    grade = examiner_grade,
    start_date,
    end_date,
    examiner_id = patex_id
  )
```

```
## Joining, by = c("examiner_name", "old_pid", "new_pid")
```

```
time_in_grade <- examiner_gs %>%
  mutate(
    start_date = mdy(start_date), # converting into proper date type
    end_date = mdy(end_date), # converting into proper date type
    days_in_grade = interval(start_date, end_date) %/% days(1)
  ) %>%
  group_by(examiner_id) %>%
  filter(grade!=max(grade, na.rm = TRUE)) %>% # dropping the highest grade record
  summarise(mean_days_in_grade = mean(days_in_grade, na.rm = TRUE))
time_in_grade
```

```
## # A tibble: 10,860 × 2
##    examiner_id mean_days_in_grade
##          <dbl>              <dbl>
##  1       59012               356.
##  2       59015               783
##  3       59016               341.
##  4       59018               368.
##  5       59019               293
##  6       59025               485
##  7       59027               364.
##  8       59030               493.
##  9       59033               258.
## 10       59035               308.
## # … with 10,850 more rows
```

```
examiner_data <- app_data_sample %>%
  filter(disposal_type!="PEND") %>% # here, we exclude in-process applications
  mutate(
    app_start_date = ymd(filing_date),
    app_end_date = case_when(
      disposal_type == "ISS" ~ ymd(patent_issue_date), # for issued patents
      disposal_type == "ABN" ~ ymd(abandon_date), # for abandoned applications
      TRUE ~ NA_Date_
    ),
    app_proc_days = interval(app_start_date, app_end_date) %/% days(1)) %>%
  filter(app_proc_days>0 & app_proc_days < 3650) %>% # limit to 0-10 years
  group_by(examiner_id) %>%
  summarise(
    app_count = n(),
    tc = min(tc, na.rm = TRUE),
    gender = first(gender),
    race = first(race),
    tenure_days = max(tenure_days, na.rm = TRUE),
    mean_app_proc_days = mean(app_proc_days, na.rm = TRUE)
  )
examiner_data
```

```
## # A tibble: 5,549 × 7
##    examiner_id app_count    tc gender race  tenure_days mean_app_proc_days
##          <dbl>     <int> <dbl> <chr>  <chr>       <dbl>              <dbl>
##  1       59012        84  1700 male   white        4013              1295.
##  2       59025         96  2400 male   Asian        2761              1152.
##  3       59030        358  2400 <NA>   black        4179              1008.
##  4       59040        233  1700 female Asian        3542              1305.
##  5       59052          8  2100 male   Asian        2017               535.
##  6       59054         10  2100 <NA>   Asian        5887              1297 
##  7       59055          2  2100 male   Asian        1149               932.
##  8       59056       1019  2100 male   Asian        6268              1077.
##  9       59074        166  2100 <NA>   white        6255              1579.
## 10       59081         48  2400 male   Asian        2220              1317.
## # … with 5,539 more rows
```

```
examiner_data <- examiner_data %>%
  left_join(time_in_grade)
```

```
## Joining, by = "examiner_id"
```

```
examiner_data
```

```
## # A tibble: 5,549 × 8
##    examiner_id app_count    tc gender race  tenure_days mean_app_proc_days
##          <dbl>     <int> <dbl> <chr>  <chr>       <dbl>              <dbl>
##  1       59012         84  1700 male   white        4013              1295.
##  2       59025         96  2400 male   Asian        2761              1152.
##  3       59030        358  2400 <NA>   black        4179              1008.
##  4       59040        233  1700 female Asian        3542              1305.
##  5       59052          8  2100 male   Asian        2017               535.
##  6       59054         10  2100 <NA>   Asian        5887              1297 
##  7       59055          2  2100 male   Asian        1149               932.
##  8       59056       1019  2100 male   Asian        6268              1077.
##  9       59074        166  2100 <NA>   white        6255              1579.
## 10       59081         48  2400 male   Asian        2220              1317.
## # … with 5,539 more rows, and 1 more variable: mean_days_in_grade <dbl>
```

```
library(modelsummary)
models <- list()
models[['m1']] <- lm(mean_days_in_grade ~ 1 + mean_app_proc_days, data = examiner_dat
models[['m2']] <- lm(mean_days_in_grade ~ 1 + mean_app_proc_days + as_factor(race),
    data = examiner_data)
models[['m3']] <- lm(mean_days_in_grade ~ 1 + mean_app_proc_days + as_factor(gender),
         data = examiner_data)
modelsummary(models)
```

	m1	m2	m3
(Intercept)	528.481	531.761	550.975
	(43.856)	(44.213)	(49.860)
mean_app_proc_days	0.014	0.016	-0.004
	(0.035)	(0.035)	(0.039)
as_factor(race)Asian		-17.130	
		(21.627)	
as_factor(race)black		38.196	
		(49.231)	
as_factor(race)Hispanic		-46.940	
		(49.354)	
as_factor(race)other		-86.266	
		(654.746)	
as_factor(gender)female			-4.166
			(23.854)
Num.Obs.	4503	4503	3838
R2	0.000	0.001	0.000
R2 Adj.	0.000	-0.001	-0.001
AIC	71176.4	71182.1	60975.0
BIC	71195.6	71227.0	61000.0
Log.Lik.	-35585.191	-35584.071	-30483.507
F	0.160	0.480	0.019
RMSE	654.48	654.60	681.30

```
women_variable <- examiner_data %>%
filter(gender == "female")
mean(women_variable$mean_days_in_grade,na.rm=TRUE)
```

```
## [1] 542.1556
```

```
men_variable <- examiner_data %>%
filter(gender == "male")
mean(men_variable$mean_days_in_grade,na.rm=TRUE)
```

```
## [1] 546.1771
```

```
white_variable <- examiner_data %>%
  filter(race == "white")
mean(white_variable$mean_days_in_grade,na.rm=TRUE)
```

```
## [1] 551.0042
```

```
asian_variable <- examiner_data %>%
  filter(race == "Asian")
mean(asian_variable$mean_days_in_grade,na.rm=TRUE)
```

```
## [1] 534.5086
```

```
black_variable <- examiner_data %>%
  filter(race == "black")
mean(black_variable$mean_days_in_grade,na.rm=TRUE)
```

```
## [1] 589.2735
```

```
hispanic_variable <- examiner_data %>%
  filter(race == "Hispanic")
mean(hispanic_variable$mean_days_in_grade,na.rm=TRUE)
```

```
## [1] 504.2059
```

There seems to be no real difference for Gender when it comes to promotion. When it comes to race They also seem closely related, however black people tend to take longest at 589 days and hispanic the least at 504 days.

From the means and regression model summary, there seems to not be any effect of gender on race on the time it takes to get a promotion.

There could be some limitations such as: Not standardized method of promoting people. Different examiners or promoters could have different biases. Other Factors such as work ethic/production that are better indicators of promotion. Assumes they all do the same job or department. Different departments could have different criteria or dominant race/gender combos.