Exercise 4

Adityo Das Gupta

03/04/2023

1. Load data, get gender, and create `app_proc_time` column

Load data

```
data_path <- "/Users/adityodasgupta/Documents/McGill/ORGB/672_project_data/"
applications <- read_parquet(paste0(data_path,"app_data_sample.parquet"))
edges <- read_csv(paste0(data_path,"edges_sample.csv"))
```

```
## 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.
```

```
applications
```

application_number <chr>	filing_date <date>	examiner_name_last <chr>	examiner_name_first <chr>	examiner_name_middle <chr>
08284457	2000-01-26	HOWARD	JACQUELINE	V
08413193	2000-10-11	YILDIRIM	BEKIR	L
08531853	2000-05-17	HAMILTON	CYNTHIA	NA
08637752	2001-07-20	MOSHER	MARY	NA
08682726	2000-04-10	BARR	MICHAEL	E
08687412	2000-04-28	GRAY	LINDA	LAMEY
08716371	2004-01-26	MCMILLIAN	KARA	RENITA
08765941	2000-06-23	FORD	VANESSA	L
08776818	2000-02-04	STRZELECKA	TERESA	E
08809677	2002-02-20	KIM	SUN	U

1-10 of 10,000 rows | 1-5 of 16 columns Previous 1 2 3 4 5 6 ... 1000 Next

```
edges
```

application_number <chr>	advice_date <date>	ego_examiner_id <dbl>	alter_examiner_id <dbl>
09402488	2008-11-17	84356	66266
09402488	2008-11-17	84356	63519
09402488	2008-11-17	84356	98531
09445135	2008-08-21	92953	71313
09445135	2008-08-21	92953	93865
09445135	2008-08-21	92953	91818
09479304	2008-12-15	61767	69277
09479304	2008-12-15	61767	92446
09479304	2008-12-15	61767	66805
09479304	2008-12-15	61767	70919

1-10 of 10,000 rows Previous 1 2 3 4 5 6 ... 1000 Next

Get gender for examiners

```
#install_genderdata_package() # only run this line the first time you use the package, to get data for it
# get a list of first names without repetitions
examiner_names <- applications %>%
  distinct(examiner_name_first)
#Now let's use function `gender()` as shown in the example for the package to attach a gender and probability to
each name and put the results into the table `examiner_names_gender`
# get a table of names and gender
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
  )
examiner_names_gender
```

examiner_name_first <chr>	gender <chr>	proportion_female <dbl>
AARON	male	0.0082
ABDEL	male	0.0000
ABDOU	male	0.0000
ABDUL	male	0.0000
ABDULHAKIM	male	0.0000
ABDULLAH	male	0.0000
ABDULLAHI	male	0.0000
ABIGAIL	female	0.9982
ABIMBOLA	female	0.9436
ABRAHAM	male	0.0031

1-10 of 1,822 rows Previous 1 2 3 4 5 6 ... 183 Next

```
# Finally, let's join that table back to our original applications data and discard the temporary tables we have
just created to reduce clutter in our environment.
# remove extra columns from the gender table
examiner_names_gender <- examiner_names_gender %>%
  select(examiner_name_first, gender)
# joining gender back to the dataset
applications <- applications %>%
  left_join(examiner_names_gender, by = "examiner_name_first")
# cleaning up
rm(examiner_names)
rm(examiner_names_gender)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells  4840433 258.6    8092220 432.2         NA  5198797 277.7
## Vcells 50504360 385.4   96616289 737.2      16384 80820144 616.7
```

Guess the examiner's race

```
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct()
examiner_surnames
```

surname <chr>
HOWARD
YILDIRIM
HAMILTON
MOSHER
BARR
GRAY
MCMILLIAN
FORD
STRZELECKA
KIM

1-10 of 3,806 rows Previous 1 2 3 4 5 6 ... 381 Next

```
examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>%
  as_tibble()
```

```
## Warning: Unknown or uninitialised column: `state`.
```

```
## Proceeding with last name predictions...
```

```
## ℹ All local files already up-to-date!
```

```
## 701 (18.4%) individuals' last names were not matched.
```

```
examiner_race
```

surname <chr>	pred.whi <dbl>	pred.bla <dbl>	pred.his <dbl>	pred.asi <dbl>	pred.oth <dbl>
HOWARD	0.596663827	0.2948321643	0.0275207004	0.0068983230	0.074084985
YILDIRIM	0.806879507	0.0273312368	0.0694453842	0.0165026000	0.079841272
HAMILTON	0.655944140	0.2386940681	0.0286452612	0.0074999249	0.069216606
MOSHER	0.914757789	0.0042516658	0.0290908584	0.0091684132	0.042731274
BARR	0.783554996	0.1198565790	0.0267955895	0.0083027220	0.061490113
GRAY	0.639787385	0.2522555912	0.0280982391	0.0074809225	0.072377863
MCMILLIAN	0.321548611	0.5544059586	0.0211825127	0.0034011419	0.099461776
FORD	0.576284887	0.3202528684	0.0275472214	0.0062098871	0.069705137
STRZELECKA	0.472364190	0.1708059537	0.2200066888	0.0825296368	0.054293531
KIM	0.016912615	0.0028201203	0.0054594943	0.9429391851	0.031868586

1-10 of 3,806 rows Previous 1 2 3 4 5 6 ... 381 Next

```
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
```

surname <chr>	pred.whi <dbl>	pred.bla <dbl>	pred.his <dbl>	pred.asi <dbl>	pred.oth <dbl>	max_race_p <dbl>	race <chr>
HOWARD	0.596663827	0.2948321643	0.0275207004	0.0068983230	0.074084985	0.5966638	white
YILDIRIM	0.806879507	0.0273312368	0.0694453842	0.0165026000	0.079841272	0.8068795	white
HAMILTON	0.655944140	0.2386940681	0.0286452612	0.0074999249	0.069216606	0.6559441	white
MOSHER	0.914757789	0.0042516658	0.0290908584	0.0091684132	0.042731274	0.9147578	white
BARR	0.783554996	0.1198565790	0.0267955895	0.0083027220	0.061490113	0.7835550	white
GRAY	0.639787385	0.2522555912	0.0280982391	0.0074809225	0.072377863	0.6397874	white
MCMILLIAN	0.321548611	0.5544059586	0.0211825127	0.0034011419	0.099461776	0.5544060	black
FORD	0.576284887	0.3202528684	0.0275472214	0.0062098871	0.069705137	0.5762849	white
STRZELECKA	0.472364190	0.1708059537	0.2200066888	0.0825296368	0.054293531	0.4723642	white
KIM	0.016912615	0.0028201203	0.0054594943	0.9429391851	0.031868586	0.9429392	Asian

1-10 of 3,806 rows Previous 1 2 3 4 5 6 ... 381 Next

```
# Let's join the data back to the applications table.
# removing extra columns
examiner_race <- examiner_race %>%
  select(surname,race)
applications <- applications %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))
rm(examiner_race)
rm(examiner_surnames)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
## Ncells  4910516 262.3    8092220 432.2         NA  7553737 403.5
## Vcells 54895477 418.9   96616289 737.2      16384 96616180 737.2
```

Examiner's tenure

```
examiner_dates <- applications %>%
  select(examiner_id, filing_date, appl_status_date)
examiner_dates
```

examiner_id <dbl>	filing_date <date>	appl_status_date <chr>
96082	2000-01-26	30jan2003 00:00:00
87678	2000-10-11	27sep2010 00:00:00
63213	2000-05-17	30mar2009 00:00:00
73788	2001-07-20	07sep2009 00:00:00
77294	2000-04-10	19apr2001 00:00:00
68606	2000-04-28	16jul2001 00:00:00
89557	2004-01-26	15may2017 00:00:00
97543	2000-06-23	03apr2002 00:00:00
98714	2000-02-04	27nov2002 00:00:00
65530	2002-02-20	23mar2009 00:00:00

1-10 of 10,000 rows Previous 1 2 3 4 5 6 ... 1000 Next

```
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)
examiner_dates
```

examiner_id <dbl>	earliest_date <date>	latest_date <date>	tenure_days <dbl>
59012	2004-07-28	2015-07-24	4013
59025	2009-10-26	2017-05-18	2761
59030	2005-12-12	2017-05-22	4179
59040	2007-09-11	2017-05-23	3542
59052	2001-08-21	2007-02-28	2017
59054	2000-11-10	2016-12-23	5887
59055	2004-11-02	2007-12-26	1149
59056	2000-03-24	2017-05-22	6268
59074	2000-01-31	2017-03-17	6255
59081	2011-04-21	2017-05-19	2220

1-10 of 5,625 rows Previous 1 2 3 4 5 6 ... 563 Next

```
# Joining back to the applications data.
applications <- applications %>%
  left_join(examiner_dates, by = "examiner_id")
rm(examiner_dates)
gc()
```

```
##            used  (Mb) gc trigger   (Mb) limit (Mb)  max used   (Mb)
## Ncells  4916940 262.6   14380074  768.0         NA  14380074  768.0
## Vcells 65271493 498.0  139303455 1062.9      16384 139303176 1062.8
```

Application Processing Time

Clean Data

```
# Remove Nas from status date and gender
applications <- applications %>%
  filter(!is.na(appl_status_date) | !is.na(gender) | !is.na(race))
# Clean Date format
#get the date format cleaned
applications$Date_time=as.Date(applications$appl_status_date, format="%d%b%Y")
#get the date format for the filing date cleaned
applications$filing_date=as.Date(applications$filing_date, format="%d%b%Y")
```

Pre-process

```
#Remove all the data we will not need based on application status
exclude_list=c("PEND")
applications <- applications %>%
  filter(!disposal_type %in% exclude_list)
```

```
#Setting Gender as factor
applications$gender = as.factor(applications$gender)
#Setting ethnicity as factor
applications$race = as.factor(applications$race)
#Setting disposal type as factor
applications$disposal_type = as.factor(applications$disposal_type)
#setting the technology center as a factor
applications$tc = as.factor(applications$tc)
```

1. Create 'app_proc_time'

```
#this is the amount of time in days that the applications take
applications$app_proc_time <- applications$Date_time - applications$filing_date
applications$app_proc_time <- as.numeric(applications$app_proc_time)
```

##Nodes & Edges Creation First we need to create the netwrok data to calculate centrality We will remove any records that contain NAs to avoid future issues with coding

```
#Create the edges from edge data
edges_backup=edges
#edges=edges_backup
edges <- edges %>%
  mutate(from=ego_examiner_id,to=alter_examiner_id) %>%
  select(from, to) %>%
  drop_na()
#Create Nodes from Edges Data
nodes <-as.data.frame(do.call(rbind,append(as.list(edges$from),as.list(edges$to))))
nodes <- nodes %>%
  mutate(id=V1) %>%
  select(id) %>%
  distinct(id) %>%
  drop_na()
```

Closeness Measures

We will now add 3 closeness measures to the nodes data frame:

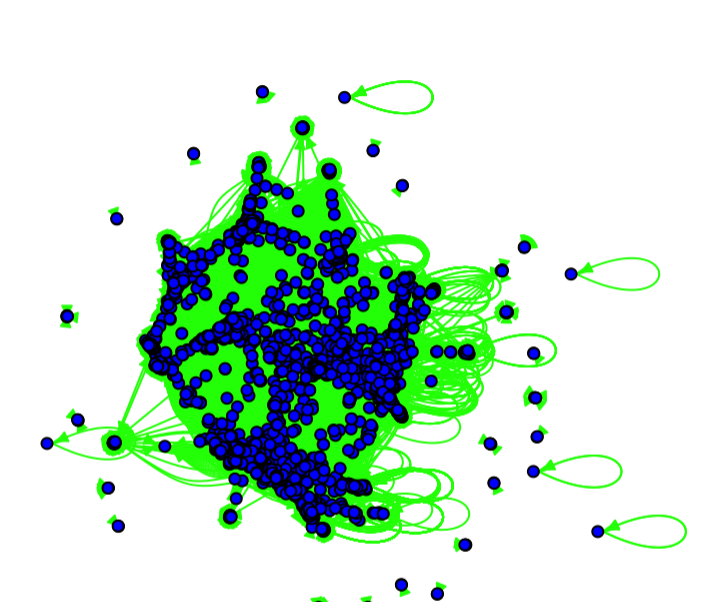
1.Degree Centrality: The number of connections (or edges) that each node has. 2. Closness Centrality : A measure that calculates the ability to spread information efficiently via the edges the node is connected to. It is calculated as the inverse of the average shortest path between nodes. 3: Betweenness Centrality: A measure that detects a node's influence over the flow of information within a graph.

```
g <- igraph::graph_from_data_frame(edges, vertices = nodes) %>% as_tbl_graph(directed=TRUE)
#not sure why this isnt working
#g = tbl_graph(nodes = nodes, edges = edges, directed = FALSE)
g <- g %>%
  activate(nodes) %>%
  mutate(degree_cen = centrality_degree(),
         closeness_cen = centrality_closeness(),
         betweenness_cen = centrality_betweenness()) %>%
  activate(edges)
tg_nodes <-
  g %>%
  activate(nodes) %>%
  data.frame() %>%
  mutate(name=as.integer(name))
nodes <- nodes %>%
  left_join(tg_nodes,by=c("id"="name"))
remove(g,tg_nodes)
```

Time to visualise the degree centralities and numeric data

```
final_data <- applications %>%
  left_join(nodes,by=c("examiner_id"="id"))

net <- igraph::graph_from_data_frame(edges, vertices = nodes) %>% as_tbl_graph(directed=TRUE)
plot(net, edge.arrow.size=.4,vertex.label=NA,vertex.size=4,vertex.color="blue",
     edge.color="green")
```



```
# Degree centrality linear regression model
model_degree <- lm(app_proc_time ~ degree_cen + gender + race + tenure_days, data = final_data)

# Betweenness centrality linear regression model
model_betweenness <- lm(app_proc_time ~ betweenness_cen + gender + race + tenure_days, data = final_data)

# Closeness centrality linear regression model
model_closeness <- lm(app_proc_time ~ closeness_cen + gender + race + tenure_days, data = final_data)

# Display the model summaries
summary(model_degree)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree_cen + gender + race + tenure_days,
##     data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1575.7  -662.8  -280.2   327.6  4727.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  777.94239    9.70831  80.132  < 2e-16 ***
## degree_cen    -0.30643    0.04044  -7.578 3.51e-14 ***
## gendermale    -4.01200    2.32959  -1.722 0.085035 .
## raceblack    -56.75960    6.00087  -9.459  < 2e-16 ***
## raceHispanic  64.06730    7.64724   8.378  < 2e-16 ***
## raceother    122.21968   31.55291   3.873 0.000107 ***
## racewhite     -6.13950    2.43558  -2.521 0.011710 *
## tenure_days    0.12773    0.00157  81.357  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1017 on 906469 degrees of freedom
##   (782240 observations deleted due to missingness)
## Multiple R-squared:  0.007769,   Adjusted R-squared:  0.007761
## F-statistic:  1014 on 7 and 906469 DF,  p-value: < 2.2e-16
```

```
summary(model_betweenness)
```

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness_cen + gender + race +
##     tenure_days, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1767.4  -662.3  -280.1   326.8  4723.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.713e+02  9.616e+00  80.217  < 2e-16 ***
## betweenness_cen  2.740e-03  1.631e-04  16.797  < 2e-16 ***
## gendermale      -4.854e+00  2.330e+00  -2.083   0.0372 *
## raceblack       -5.345e+01  5.999e+00  -8.911  < 2e-16 ***
## raceHispanic     6.729e+01  7.648e+00   8.799  < 2e-16 ***
## raceother        1.273e+02  3.155e+01   4.036 5.43e-05 ***
## racewhite       -4.723e+00  2.435e+00  -1.939   0.0525 .
## tenure_days      1.279e-01  1.564e-03  81.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1017 on 906469 degrees of freedom
##   (782240 observations deleted due to missingness)
## Multiple R-squared:  0.008015,   Adjusted R-squared:  0.008007
## F-statistic:  1046 on 7 and 906469 DF,  p-value: < 2.2e-16
```

```
summary(model_closeness)
```

```
##
## Call:
## lm(formula = app_proc_time ~ closeness_cen + gender + race +
##     tenure_days, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1551.0  -634.6  -252.3   338.0  4739.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   840.624051  12.013499  69.973  < 2e-16 ***
## closeness_cen -54.268985   3.522769 -15.405  < 2e-16 ***
## gendermale     -7.600628   2.738096  -2.776  0.00551 **
## raceblack      -8.696138   7.157424  -1.215  0.22437
## raceHispanic   26.742143   8.680627   3.081  0.00207 **
## raceother      16.020405  54.142223   0.296  0.76731
## racewhite     -35.515944   2.904359 -12.228  < 2e-16 ***
## tenure_days     0.122323   0.001984  61.646  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 968.3 on 588079 degrees of freedom
##   (1100630 observations deleted due to missingness)
## Multiple R-squared:  0.006878,   Adjusted R-squared:  0.006867
## F-statistic: 581.9 on 7 and 588079 DF,  p-value: < 2.2e-16
```

Get the summary of the linear regressions!

```
model_degree_interaction <- lm(app_proc_time ~ degree_cen * gender + race + tenure_days, data = final_data)
model_betweenness_interaction <- lm(app_proc_time ~ betweenness_cen * gender + race + tenure_days, data = final_d
ata)
model_closeness_interaction <- lm(app_proc_time ~ closeness_cen * gender + race + tenure_days, data = final_data)

summary(model_degree_interaction)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree_cen * gender + race + tenure_days,
##     data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1573.7  -662.8  -280.1   327.6  4728.3
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             773.56395    9.74887  79.349  < 2e-16 ***
## degree_cen                0.05449    0.08375   0.651 0.515280
## gendermale                0.61030    2.51179   0.243 0.808026
## raceblack               -56.70939    6.00081  -9.450  < 2e-16 ***
## raceHispanic             64.93067    7.64916   8.489  < 2e-16 ***
## raceother               121.68636   31.55269   3.857 0.000115 ***
## racewhite                -6.27995    2.43571  -2.578 0.009929 **
## tenure_days               0.12788    0.00157  81.437  < 2e-16 ***
## degree_cen:gendermale    -0.46938    0.09538  -4.921 8.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1017 on 906468 degrees of freedom
##   (782240 observations deleted due to missingness)
## Multiple R-squared:  0.007795,   Adjusted R-squared:  0.007786
## F-statistic: 890.2 on 8 and 906468 DF,  p-value: < 2.2e-16
```

```
summary(model_betweenness_interaction)
```

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness_cen * gender + race +
##     tenure_days, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1823.4  -662.3  -280.0   326.7  4727.5
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 7.721e+02  9.616e+00  80.289  < 2e-16 ***
## betweenness_cen             9.875e-04  3.439e-04   2.871  0.00409 **
## gendermale                 -6.493e+00  2.347e+00  -2.766  0.00567 **
## raceblack                  -5.360e+01  5.999e+00  -8.934  < 2e-16 ***
## raceHispanic                6.756e+01  7.648e+00   8.834  < 2e-16 ***
## raceother                   1.281e+02  3.155e+01   4.062 4.87e-05 ***
## racewhite                  -4.291e+00  2.436e+00  -1.761  0.07816 .
## tenure_days                 1.279e-01  1.564e-03  81.766  < 2e-16 ***
## betweenness_cen:gendermale  2.261e-03  3.906e-04   5.789 7.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1017 on 906468 degrees of freedom
##   (782240 observations deleted due to missingness)
## Multiple R-squared:  0.008051,   Adjusted R-squared:  0.008043
## F-statistic: 919.7 on 8 and 906468 DF,  p-value: < 2.2e-16
```

```
summary(model_closeness_interaction)
```

```
##
## Call:
## lm(formula = app_proc_time ~ closeness_cen * gender + race +
##     tenure_days, data = final_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1579.5  -634.6  -252.5   338.2  4739.2
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              823.204934  12.207231  67.436  < 2e-16 ***
## closeness_cen            -13.385628   6.192129  -2.162  0.03064 *
## gendermale                11.433802   3.621863   3.157  0.00159 **
## raceblack                -12.332807   7.171360  -1.720  0.08548 .
## raceHispanic              21.047666   8.709093   2.417  0.01566 *
## raceother                 11.174069  54.142668   0.206  0.83649
## racewhite                -35.836249   2.904477 -12.338  < 2e-16 ***
## tenure_days                0.123109   0.001987  61.970  < 2e-16 ***
## closeness_cen:gendermale -59.756536   7.443424  -8.028 9.92e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 968.2 on 588078 degrees of freedom
##   (1100630 observations deleted due to missingness)
## Multiple R-squared:  0.006987,   Adjusted R-squared:  0.006974
## F-statistic: 517.2 on 8 and 588078 DF,  p-value: < 2.2e-16
```

Interpretations:

On an average looking at the linear regression models:

if the race is white application processing time decreases by the most

if the race is hispanic application processing time increases by the most

if gender is male it takes less time than female

longer the tenure more the time taken

and if a male is processing an application of another male then it makes a significant decrease in time