Exercise 3

2024-03-31

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
```

```
applications <- read_parquet("C:/Users/xzhu71/Desktop/app_data_sample.parquet")
edges <- read_csv(paste0("C:/Users/xzhu71/Desktop/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.
```

add gender variables for examiners

```
#install_genderdata_package()
# get a list of first names without repetitions
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,
    proportion_female
  )

examiner_names_gender
```

```
## # A tibble: 1,822 x 3
##    examiner_name_first gender proportion_female
##    <chr>               <chr>              <dbl>
##  1 AARON               male              0.0082
##  2 ABDEL               male              0
##  3 ABDOU               male              0
##  4 ABDUL               male              0
##  5 ABDULHAKIM          male              0
##  6 ABDULLAH            male              0
##  7 ABDULLAHI           male              0
##  8 ABIGAIL             female            0.998
##  9 ABIMBOLA            female            0.944
## 10 ABRAHAM             male              0.0031
## # i 1,812 more rows
```

```r
# remove extra colums 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) max used  (Mb)
## Ncells  4748431 253.6    8323141 444.6  4768677 254.7
## Vcells 50038634 381.8   96057496 732.9 80354708 613.1
```

add race variable for the examiners

```r
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct()
examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>%
  as_tibble()
```

```
## Predicting race for 2020

## 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.
```

```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_
  ))

examiner_race
```

```
## # A tibble: 3,806 x 8
##    surname    pred.whi pred.bla pred.his pred.asi pred.oth max_race_p race
##    <chr>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>      <dbl> <chr>
##  1 HOWARD       0.597   0.295    0.0275   0.00690   0.0741      0.597 white
##  2 YILDIRIM     0.807   0.0273   0.0694   0.0165    0.0798      0.807 white
##  3 HAMILTON     0.656   0.239    0.0286   0.00750   0.0692      0.656 white
##  4 MOSHER       0.915   0.00425  0.0291   0.00917   0.0427      0.915 white
##  5 BARR         0.784   0.120    0.0268   0.00830   0.0615      0.784 white
##  6 GRAY         0.640   0.252    0.0281   0.00748   0.0724      0.640 white
##  7 MCMILLIAN    0.322   0.554    0.0212   0.00340   0.0995      0.554 black
##  8 FORD         0.576   0.320    0.0275   0.00621   0.0697      0.576 white
##  9 STRZELECKA   0.472   0.171    0.220    0.0825    0.0543      0.472 white
## 10 KIM          0.0169  0.00282  0.00546  0.943     0.0319      0.943 Asian
## # i 3,796 more rows
```

```r
# 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) max used (Mb)
## Ncells  4839543 258.5    8323141 444.6  6496074  347
## Vcells 52224708 398.5   96057496 732.9 95539631  729
```

add tenure variable in examiners

```r
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)))
```

```r
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
```

```
## # A tibble: 5,625 x 4
##    examiner_id earliest_date latest_date tenure_days
##          <dbl> <date>        <date>            <dbl>
##  1       59012 2004-07-28    2015-07-24         4013
##  2       59025 2009-10-26    2017-05-18         2761
##  3       59030 2005-12-12    2017-05-22         4179
##  4       59040 2007-09-11    2017-05-23         3542
##  5       59052 2001-08-21    2007-02-28         2017
##  6       59054 2000-11-10    2016-12-23         5887
##  7       59055 2004-11-02    2007-12-26         1149
##  8       59056 2000-03-24    2017-05-22         6268
##  9       59074 2000-01-31    2017-03-17         6255
## 10       59081 2011-04-21    2017-05-19         2220
## # i 5,615 more rows
```

```
applications <- applications %>%
  left_join(examiner_dates, by = "examiner_id")

rm(examiner_dates)
gc()
```

```
##            used  (Mb) gc trigger   (Mb)  max used  (Mb)
## Ncells  4848048 259.0    8323141  444.6   8323141 444.6
## Vcells 58299658 444.8  138498794 1056.7 115014722 877.5
```

Question 2

compare examiners’demographics

```
applications$workgroup <- substr(as.character(applications$examiner_art_unit), 1, 3)
```

```
filtered_data <- filter(applications, workgroup %in% c('176', '162'))
# Printing the summary of missing values for demographic variables
na_demographics <- filtered_data %>%
  select(gender, race, tenure_days) %>%
  summarise_all(~sum(is.na(.)))
na_demographics
```

```
## # A tibble: 1 x 3
##   gender  race tenure_days
##    <int> <int>       <int>
## 1  44338     0        5406
```

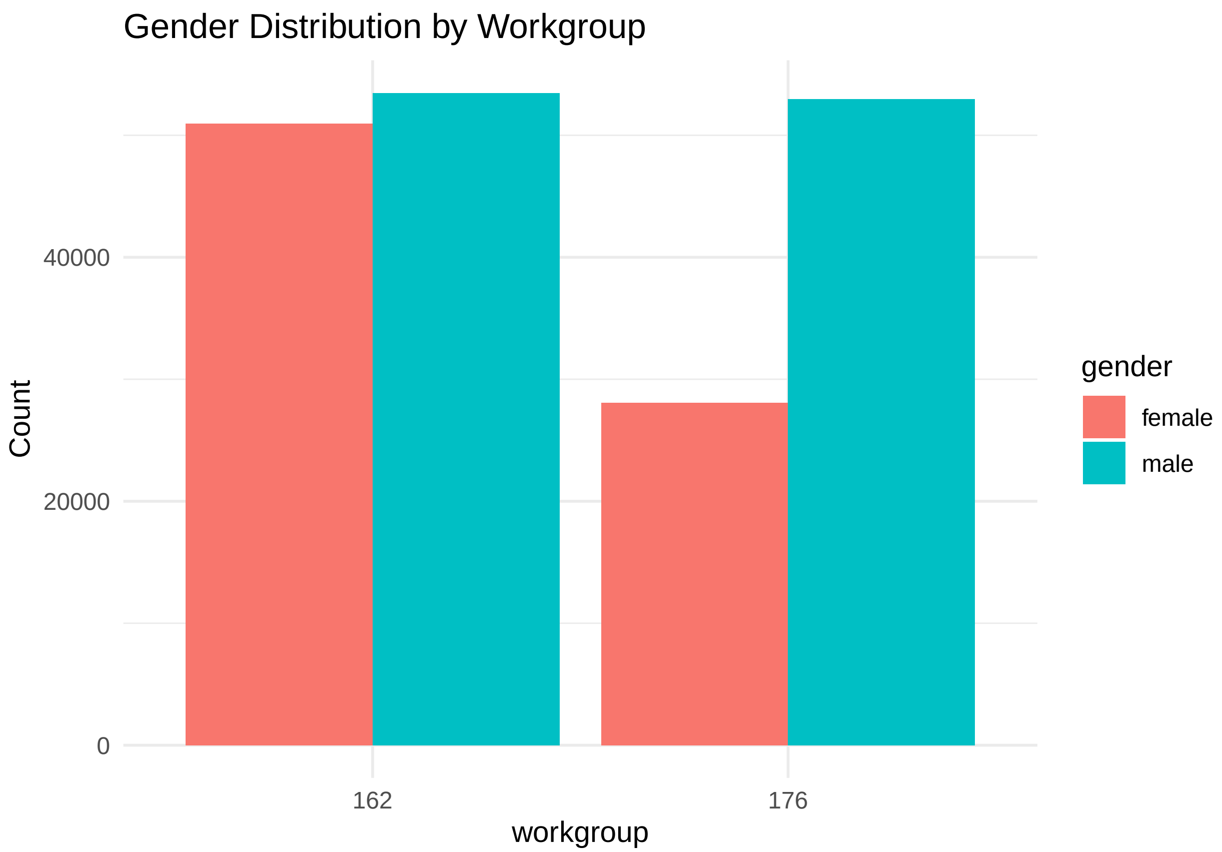
```
#drop the missing values since the portion is relatively low
filtered_data_no <- filtered_data %>%
  filter(!is.na(gender) & !is.na(tenure_days))
filtered_data_no
```

```
## # A tibble: 185,468 x 22
##    application_number filing_date examiner_name_last examiner_name_first
##    <chr>              <date>      <chr>              <chr>
##  1 08284457           2000-01-26  HOWARD             JACQUELINE
##  2 08682726           2000-04-10  BARR               MICHAEL
##  3 08716371           2004-01-26  MCMILLIAN          KARA
##  4 09068704           2001-02-23  ROBINSON           BINTA
##  5 09091481           2000-06-27  ROBINSON           BINTA
##  6 09101427           2000-09-16  BARR               MICHAEL
##  7 09125199           2000-03-23  ROTMAN             ALAN
##  8 09147568           2000-10-30  HUANG              EVELYN
##  9 09180120           2000-04-05  WONG               LESLIE
## 10 09194823           2000-02-08  RAYMOND            RICHARD
## # i 185,458 more rows
## # i 18 more variables: examiner_name_middle <chr>, examiner_id <dbl>,
## #   examiner_art_unit <dbl>, uspc_class <chr>, uspc_subclass <chr>,
## #   patent_number <chr>, patent_issue_date <date>, abandon_date <date>,
## #   disposal_type <chr>, appl_status_code <dbl>, appl_status_date <chr>,
## #   tc <dbl>, gender <chr>, race <chr>, earliest_date <date>,
## #   latest_date <date>, tenure_days <dbl>, workgroup <chr>
```

```
# Summary statistics for tenure days by workgroup
tenure_stats <- filtered_data_no %>%
  group_by(workgroup) %>%
  summarise(
    Count = n(),
    Mean = mean(tenure_days, na.rm = TRUE),
    SD = sd(tenure_days, na.rm = TRUE),
    Min = min(tenure_days, na.rm = TRUE),
    Max = max(tenure_days, na.rm = TRUE)
  )
tenure_stats
```

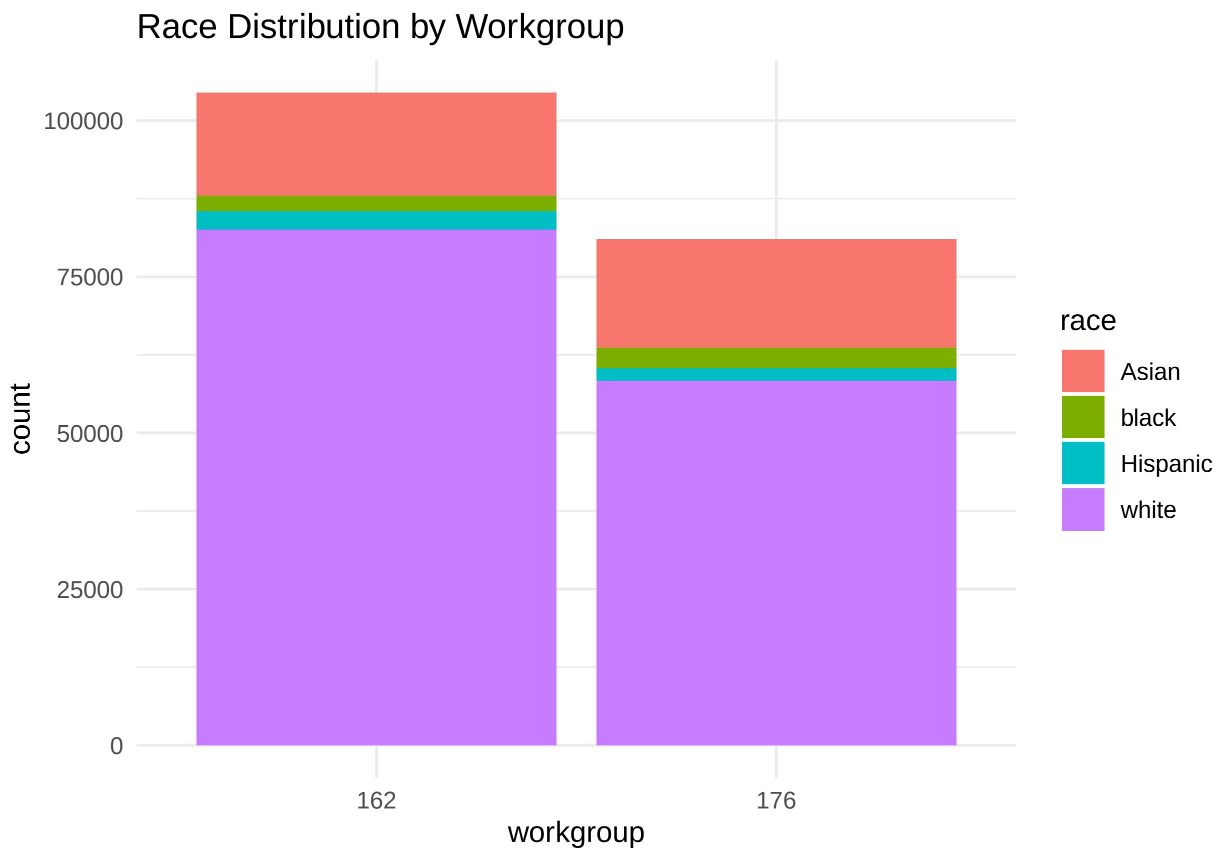
```
## # A tibble: 2 x 6
##   workgroup  Count  Mean    SD   Min   Max
##   <chr>      <int> <dbl> <dbl> <dbl> <dbl>
## 1 162       104445 5733.  768.   847  6518
## 2 176        81023 5510. 1089.   339  6350
```

```
gender_distribution <- filtered_data_no %>%
  group_by(workgroup, gender) %>%
  summarise(Count = n(), .groups = "drop") %>%
  ggplot(aes(x = workgroup, y = Count, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  ggtitle("Gender Distribution by Workgroup")

# Display the plot
gender_distribution
```



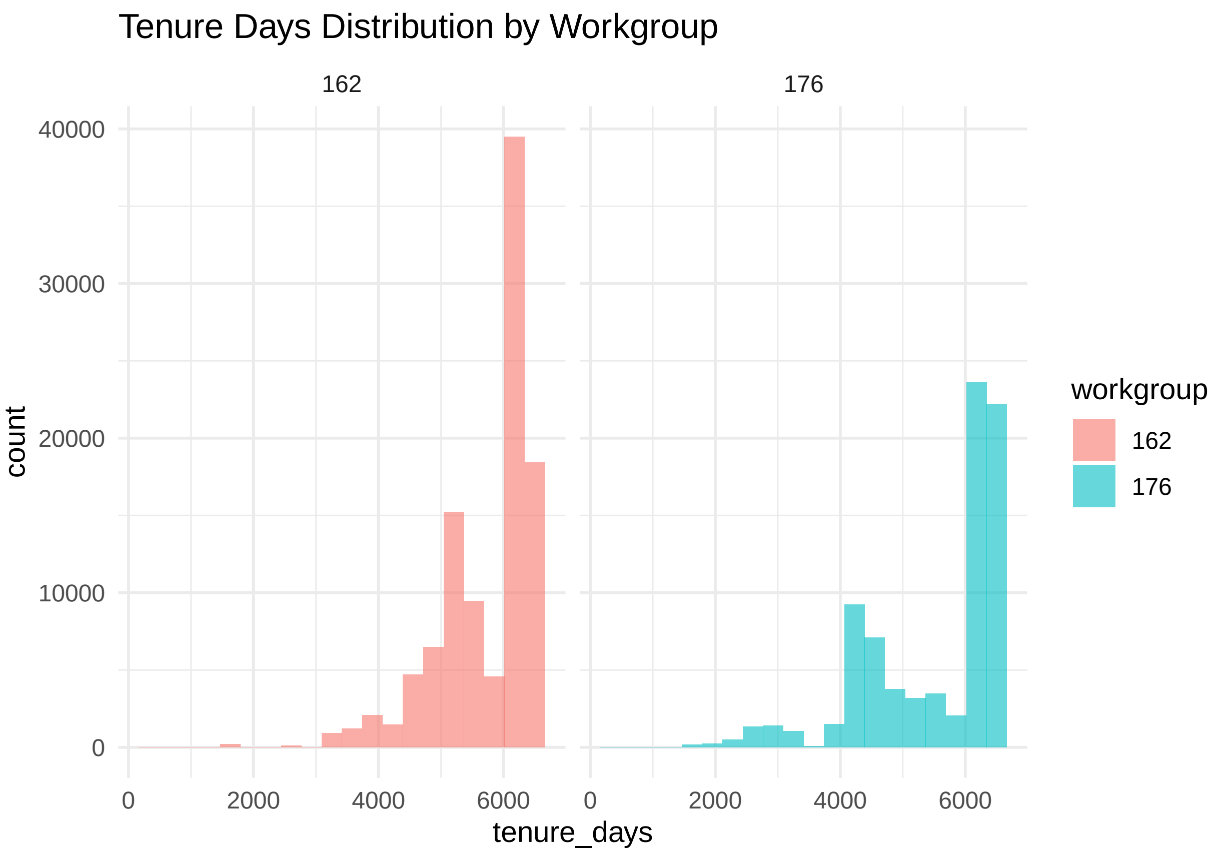
Interpretation: The bar chart illustrates the gender distribution within the two workgroups, showing a higher count of male examiners compared to female examiners in both groups.Both groups exhibit a gender imbalance favoring male examiners.

```r
# Race Distribution
ggplot(filtered_data_no, aes(x = workgroup, fill = race)) +
  geom_bar(position = "stack") +
  ggtitle("Race Distribution by Workgroup") +
  theme_minimal()
```

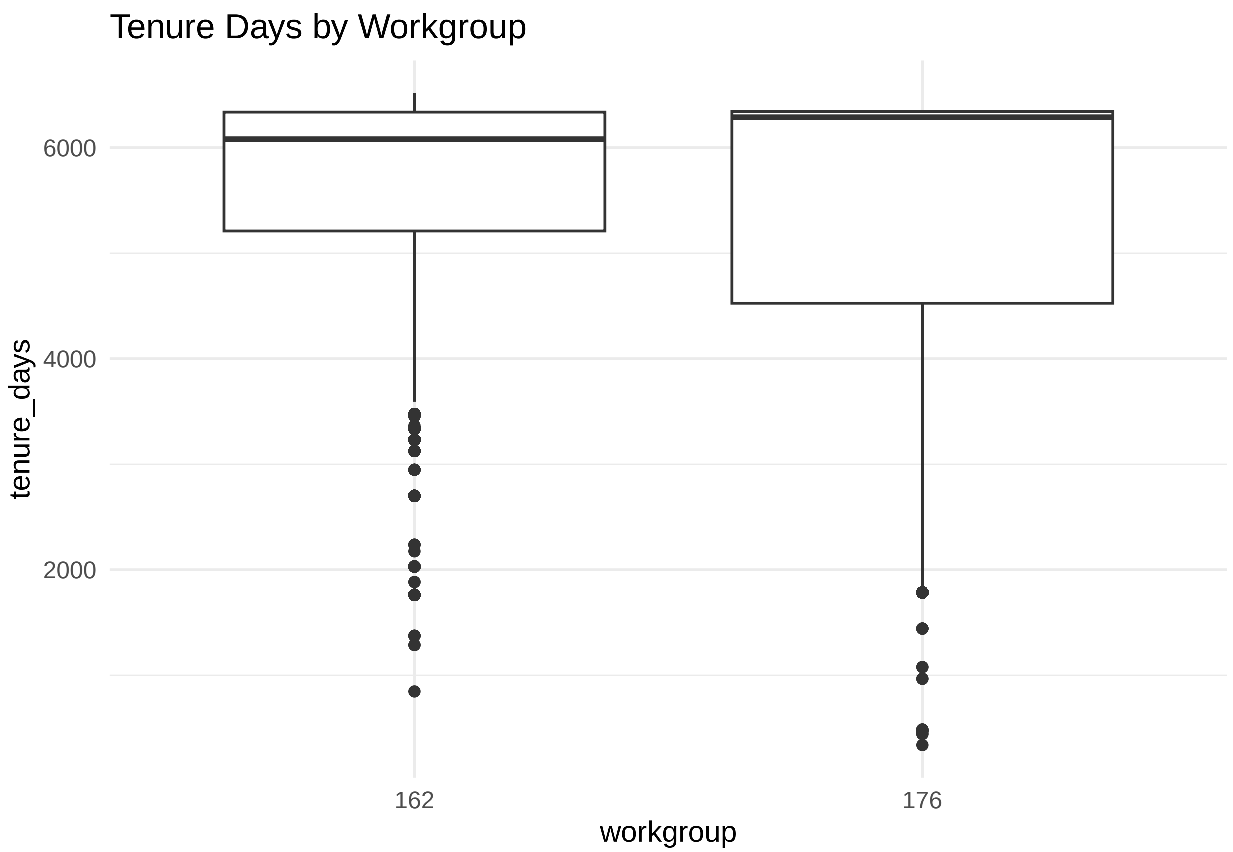


Both groups have a majority of white individuals, followed by a smaller proportion of Asian, Hispanic, and black individuals. The distribution appears relatively consistent between the two workgroups, suggesting that the racial makeup is not significantly different when comparing Workgroup 162 to Workgroup 176.

```
# Tenure Days Distribution
ggplot(filtered_data_no, aes(x = tenure_days, fill = workgroup)) +
  geom_histogram(position = "identity", alpha = 0.6, bins = 20) +
  facet_wrap(~workgroup) +
  ggtitle("Tenure Days Distribution by Workgroup") +
  theme_minimal()
```



```
# Boxplot for tenure_days across workgroups
ggplot(filtered_data_no, aes(x = workgroup, y = tenure_days)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Tenure Days by Workgroup")
```



This histogram shows the distribution of tenure days for two different workgroups, labeled 162 and 176. It appears that Workgroup 162 has a large number of individuals with a relatively short tenure, as indicated by the peak at the lower end of the tenure days axis. In contrast, Workgroup 176 shows a more uniform distribution across a range of tenure days, with a peak at the higher end. This suggests that Workgroup 176 may have more experienced examiners, or it could indicate a difference in hiring or retention patterns between the two groups.

```
joined_data <- inner_join(filtered_data_no, edges, by = "application_number")
joined_data
```

```
## # A tibble: 571 x 25
##    application_number filing_date examiner_name_last examiner_name_first
##    <chr>              <date>      <chr>              <chr>
##  1 09704054           2000-11-01  ANDERSON           JAMES
##  2 09704054           2000-11-01  ANDERSON           JAMES
##  3 09714351           2000-11-16  STOCKTON           LAURA
##  4 09714351           2000-11-16  STOCKTON           LAURA
##  5 09714351           2000-11-16  STOCKTON           LAURA
##  6 09879854           2001-06-12  MCINTOSH III       TRAVISS
##  7 09879854           2001-06-12  MCINTOSH III       TRAVISS
##  8 09879854           2001-06-12  MCINTOSH III       TRAVISS
##  9 09879854           2001-06-12  MCINTOSH III       TRAVISS
## 10 09879854           2001-06-12  MCINTOSH III       TRAVISS
## # i 561 more rows
## # i 21 more variables: examiner_name_middle <chr>, examiner_id <dbl>,
## #   examiner_art_unit <dbl>, uspc_class <chr>, uspc_subclass <chr>,
## #   patent_number <chr>, patent_issue_date <date>, abandon_date <date>,
## #   disposal_type <chr>, appl_status_code <dbl>, appl_status_date <chr>,
## #   tc <dbl>, gender <chr>, race <chr>, earliest_date <date>,
## #   latest_date <date>, tenure_days <dbl>, workgroup <chr>, ...
```

```
selected_data <- joined_data %>%
  select(application_number, examiner_name_last, examiner_name_first, examiner_id,
         gender, race, tenure_days, workgroup, ego_examiner_id, alter_examiner_id)
selected_data
```

```
## # A tibble: 571 x 10
##    application_number examiner_name_last examiner_name_first examiner_id gender
##    <chr>              <chr>              <chr>                     <dbl> <chr>
##  1 09704054           ANDERSON           JAMES                     73364 male
##  2 09704054           ANDERSON           JAMES                     73364 male
##  3 09714351           STOCKTON           LAURA                     61417 female
##  4 09714351           STOCKTON           LAURA                     61417 female
##  5 09714351           STOCKTON           LAURA                     61417 female
##  6 09879854           MCINTOSH III       TRAVISS                   67690 male
##  7 09879854           MCINTOSH III       TRAVISS                   67690 male
##  8 09879854           MCINTOSH III       TRAVISS                   67690 male
##  9 09879854           MCINTOSH III       TRAVISS                   67690 male
## 10 09879854           MCINTOSH III       TRAVISS                   67690 male
## # i 561 more rows
## # i 5 more variables: race <chr>, tenure_days <dbl>, workgroup <chr>,
## #   ego_examiner_id <dbl>, alter_examiner_id <dbl>
```

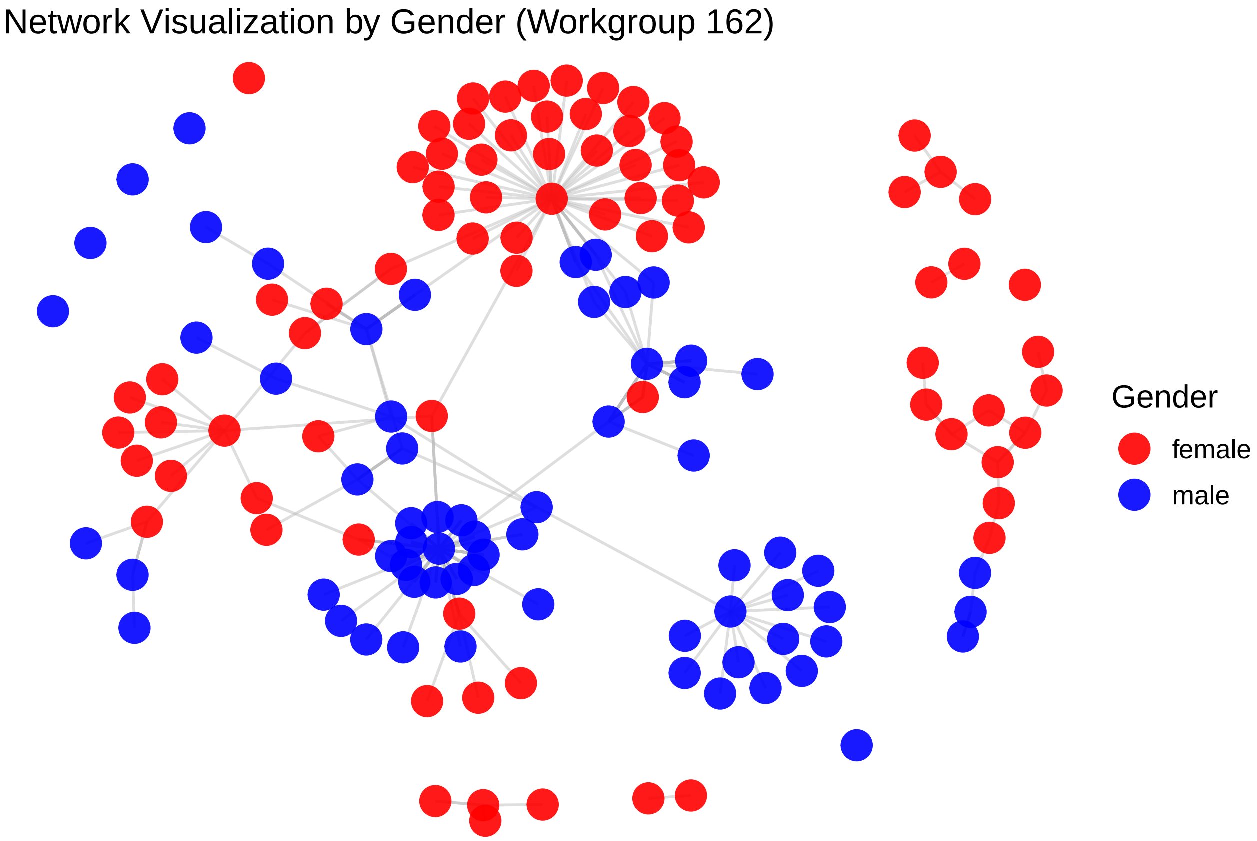
network visualization of workgroup 162

```
selected_data_162 <- filter(selected_data, workgroup == "162")
g_162 <- graph_from_data_frame(selected_data_162[, c("ego_examiner_id", "alter_examiner_id")], directed
```

```
## Warning in graph_from_data_frame(selected_data_162[, c("ego_examiner_id", : In
## ‘d’ ‘NA’ elements were replaced with string "NA"
```

```
# Combine gender information from both roles
all_genders <- rbind(
  data.frame(id = selected_data$ego_examiner_id, gender = selected_data$gender),
  data.frame(id = selected_data$alter_examiner_id, gender = selected_data$gender)
)

# Remove potential duplicates to ensure each examiner ID has a single gender entry
all_genders_unique <- all_genders[!duplicated(all_genders$id), ]
#Since some examiners might appear in both roles, this line removes duplicates to ensure that each exam

# Prepare a nodes data frame with ID from the graph
nodes_df <- data.frame(id = V(g_162)$name)
nodes_df <- merge(nodes_df, all_genders_unique, by = "id", all.x = TRUE)

# Exclude nodes with NA or "Unknown" gender
nodes_df <- nodes_df[!is.na(nodes_df$gender), ]
```

```
# Subset the graph to only include nodes with known gender information
nodes_to_keep <- as.character(nodes_df$id)
g_162_filtered <- induced_subgraph(g_162, which(V(g_162)$name %in% nodes_to_keep))
V(g_162_filtered)$gender <- nodes_df$gender[match(V(g_162_filtered)$name, nodes_df$id)]

ggraph(g_162_filtered, layout = 'fr') +
  geom_edge_link(color = 'gray', alpha = 0.5) +
  geom_node_point(aes(color = gender), size = 5, alpha = 0.9) +
  scale_color_manual(values = c('male' = 'blue', 'female' = 'red', 'Other' = 'grey')) +
  theme_void() +
  ggtitle("Network Visualization by Gender (Workgroup 162)") +
  theme(legend.title = element_text(size = 12), legend.text = element_text(size = 10)) +
  labs(color = 'Gender')
```



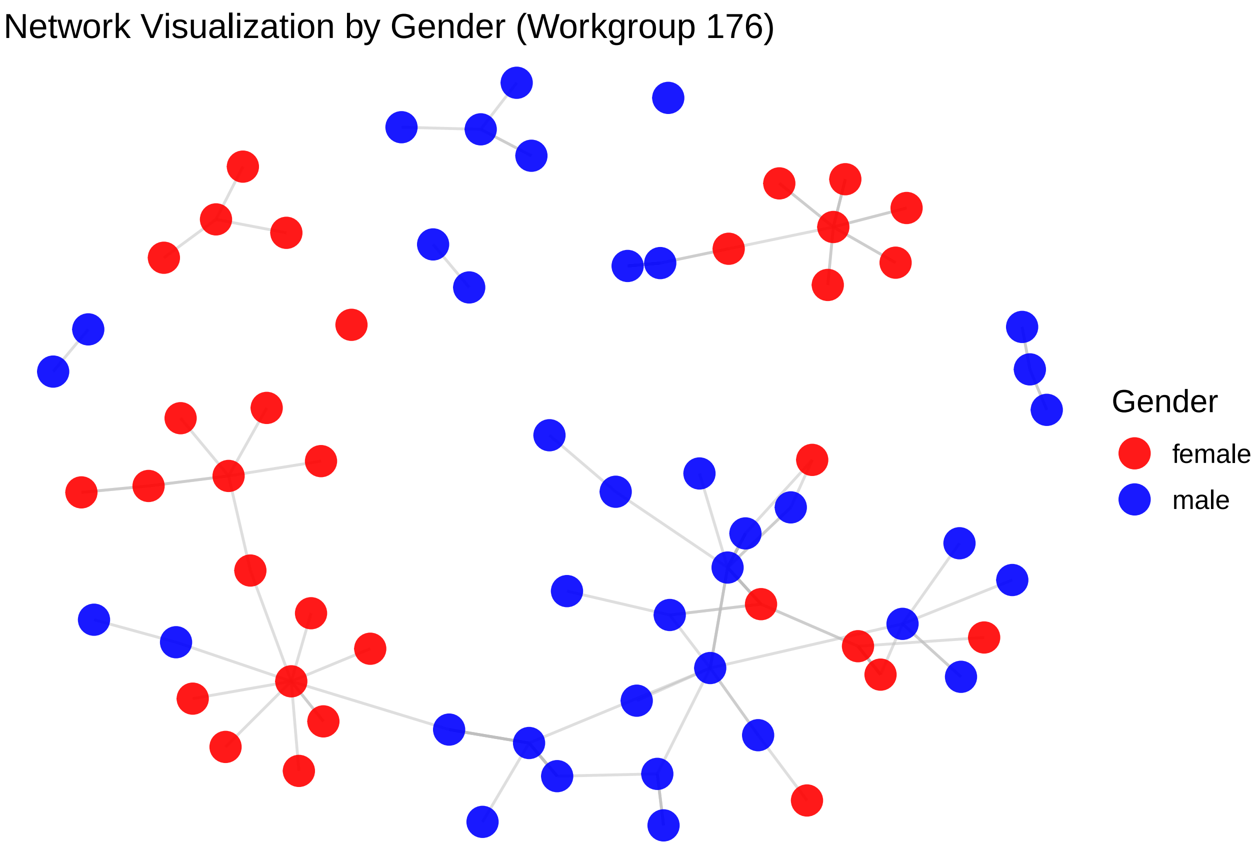
network visualization of workgroup 176

```
selected_data_176 <- filter(selected_data, workgroup == "176")
g_176 <- graph_from_data_frame(selected_data_176[, c("ego_examiner_id", "alter_examiner_id")], directed
```

```
## Warning in graph_from_data_frame(selected_data_176[, c("ego_examiner_id", : In
## ‘d’ ‘NA’ elements were replaced with string "NA"
```

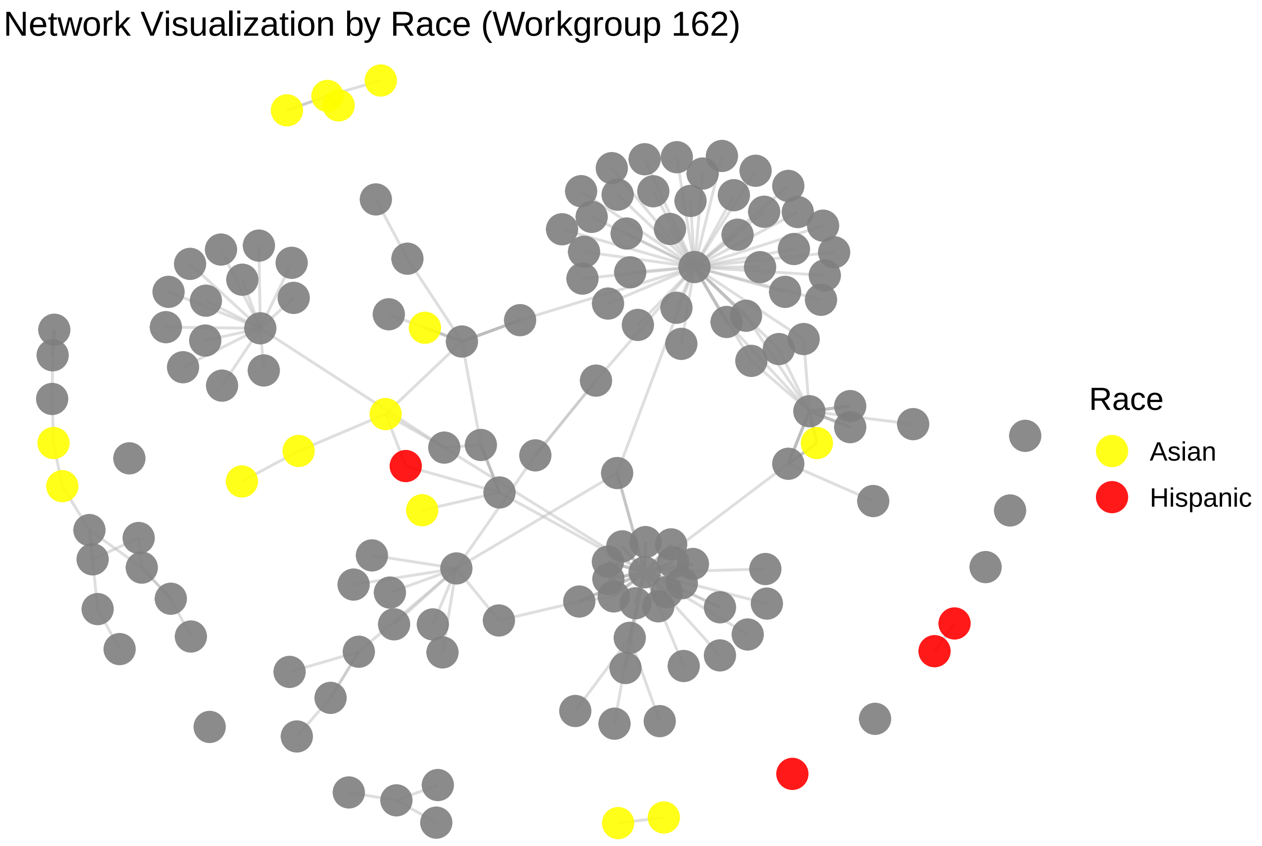
```
# Combine gender information for both roles
all_genders_176 <- rbind(
  data.frame(id = selected_data_176$ego_examiner_id, gender = selected_data_176$gender),
```

```r
  data.frame(id = selected_data_176$alter_examiner_id, gender = selected_data_176$gender)
)

# Remove potential duplicates to ensure a single gender entry per examiner ID
all_genders_unique_176 <- all_genders_176[!duplicated(all_genders_176$id), ]

# Prepare a nodes data frame with ID from the graph
nodes_df_176 <- data.frame(id = V(g_176)$name)
nodes_df_176 <- merge(nodes_df_176, all_genders_unique_176, by = "id", all.x = TRUE)

# Exclude nodes with NA gender
nodes_df_176 <- nodes_df_176[!is.na(nodes_df_176$gender), ]
nodes_to_keep_176 <- as.character(nodes_df_176$id)
g_176_filtered <- induced_subgraph(g_176, which(V(g_176)$name %in% nodes_to_keep_176))


V(g_176_filtered)$gender <- nodes_df_176$gender[match(V(g_176_filtered)$name, nodes_df_176$id)]

ggraph(g_176_filtered, layout = 'fr') +
  geom_edge_link(color = 'gray', alpha = 0.5) +
  geom_node_point(aes(color = gender), size = 5, alpha = 0.9) +
  scale_color_manual(values = c('male' = 'blue', 'female' = 'red', 'Other' = 'grey')) +
  theme_void() +
  ggtitle("Network Visualization by Gender (Workgroup 176)") +
  theme(legend.title = element_text(size = 12), legend.text = element_text(size = 10)) +
  labs(color = 'Gender')
```

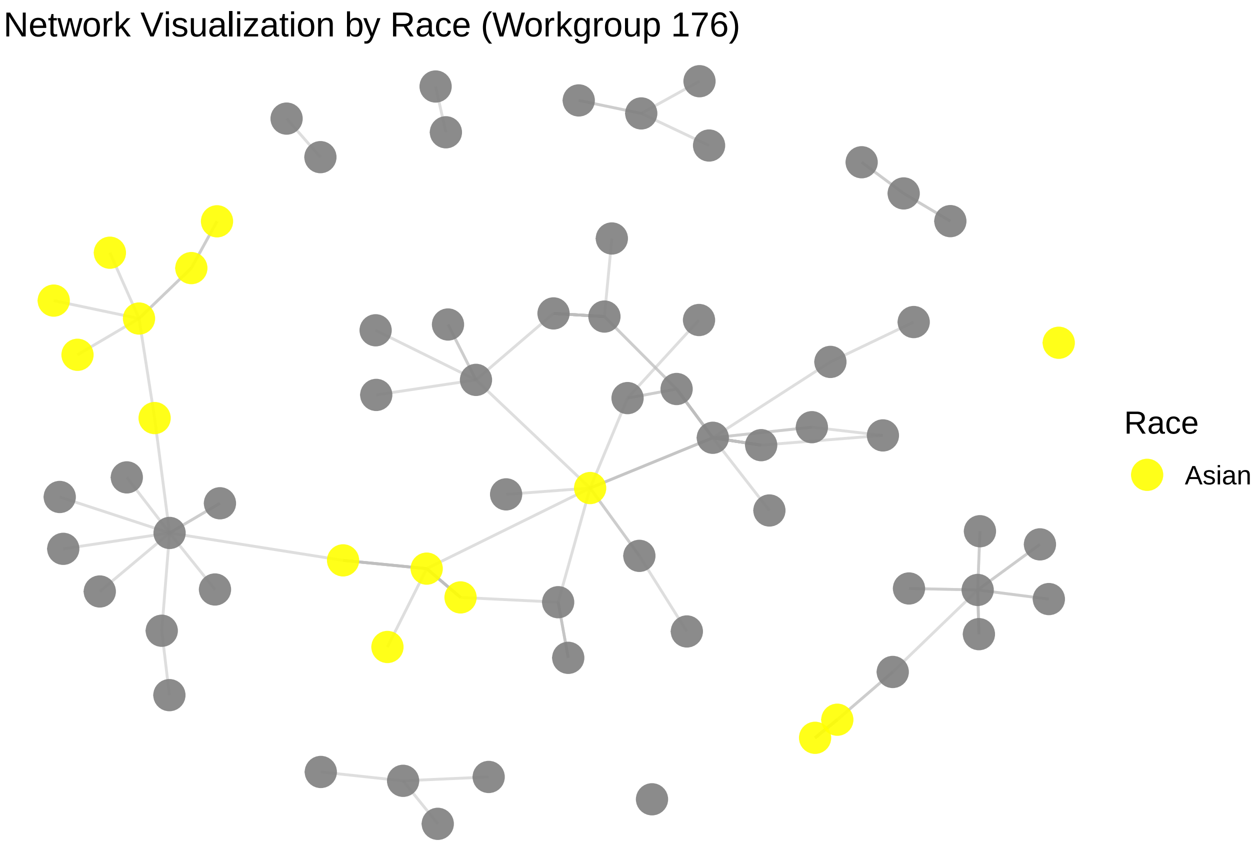


network visualization of race

```
all_races <- rbind(
  data.frame(id = selected_data$ego_examiner_id, race = selected_data$race),
  data.frame(id = selected_data$alter_examiner_id, race = selected_data$race)
)


# Remove potential duplicates to ensure a single race entry per examiner ID
all_races_unique <- all_races[!duplicated(all_races$id), ]

# Prepare a nodes data frame with ID from the graph
nodes_df <- data.frame(id = V(g_162)$name)

# Merge the nodes data frame with the consolidated race information
nodes_df <- merge(nodes_df, all_races_unique, by = "id", all.x = TRUE)

# Exclude nodes with NA race
nodes_df <- nodes_df[!is.na(nodes_df$race), ]
nodes_to_keep <- as.character(nodes_df$id)
g_162_filtered <- induced_subgraph(g_162, which(V(g_162)$name %in% nodes_to_keep))

# Update the race attribute for the filtered graph
V(g_162_filtered)$race <- nodes_df$race[match(V(g_162_filtered)$name, nodes_df$id)]

ggraph(g_162_filtered, layout = 'fr') +
  geom_edge_link(color = 'gray', alpha = 0.5) +
  geom_node_point(aes(color = race), size = 5, alpha = 0.9) +
  scale_color_manual(values = c('White' = 'blue', 'Black' = 'black', 'Asian' = 'yellow', 'Hispanic' = '
  theme_void() +
  ggtitle("Network Visualization by Race (Workgroup 162)") +
  theme(legend.title = element_text(size = 12), legend.text = element_text(size = 10)) +
  labs(color = 'Race')
```



```
nodes_df_176 <- data.frame(id = V(g_176)$name)
# Merge the nodes data frame with the consolidated race information
nodes_df_176 <- merge(nodes_df_176, all_races_unique, by = "id", all.x = TRUE)

# Exclude nodes with NA race
nodes_df_176 <- nodes_df_176[!is.na(nodes_df_176$race), ]
nodes_to_keep_176 <- as.character(nodes_df_176$id)
g_176_filtered <- induced_subgraph(g_176, which(V(g_176)$name %in% nodes_to_keep_176))

V(g_176_filtered)$race <- nodes_df_176$race[match(V(g_176_filtered)$name, nodes_df_176$id)]

# Visualization for Workgroup 176
ggraph(g_176_filtered, layout = 'fr') +
  geom_edge_link(color = 'gray', alpha = 0.5) +
  geom_node_point(aes(color = race), size = 5, alpha = 0.9) +
  scale_color_manual(values = c('White' = 'blue', 'Black' = 'black', 'Asian' = 'yellow', 'Hispanic' = '
  theme_void() +
  ggtitle("Network Visualization by Race (Workgroup 176)") +
  theme(legend.title = element_text(size = 12), legend.text = element_text(size = 10)) +
  labs(color = 'Race')
```

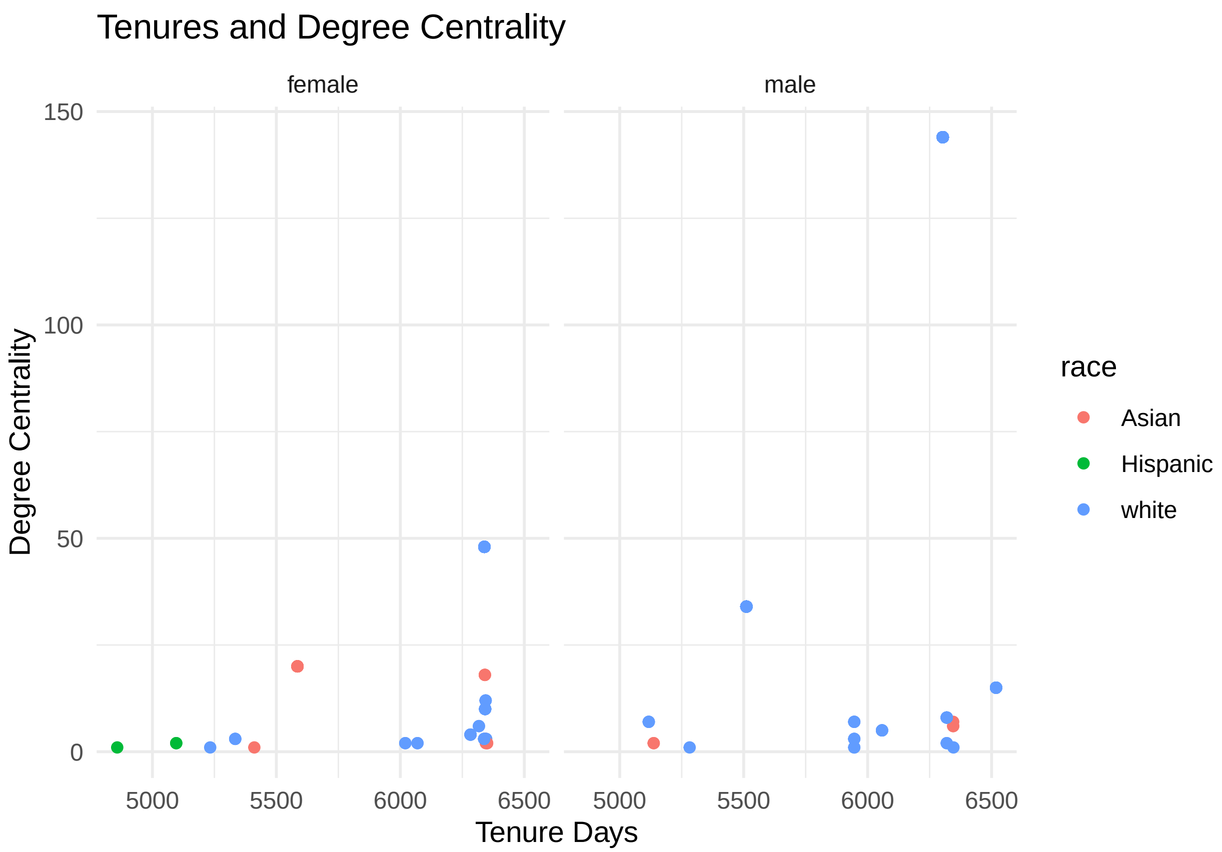


Question 3

justification for choosing degree centrality: Degree centrality was chosen because it provides a direct count of the number of connections an examiner has within the advice network, which in this context, represents the active engagement in advice-seeking (out-degree) and advice-giving (in-degree) behaviors. It is a measure of immediate influence and potential knowledge dissemination within the network, reflecting how integral an examiner is to the flow of information.

By integrating degree centrality with demographic variables such as tenure, race, and gender, we can explore the nuances of network dynamics. For instance, the code below merges centrality scores with tenure data, allowing us to analyze whether an examiner's centrality increases with their tenure.

```
degree_centrality_162 <- degree(g_162, mode = "all")
degree_centrality_176 <- degree(g_176, mode = "all")
degree_centrality_df_162 <- data.frame(ego_examiner_id = names(degree_centrality_162),
                                       degree_centrality = degree_centrality_162)

selected_data_162 <- merge(selected_data_162, degree_centrality_df_162, by = "ego_examiner_id")



ggplot(selected_data_162, aes(x = tenure_days, y = degree_centrality, color = race)) +
  geom_point() +
  facet_wrap(~gender) +
  theme_minimal() +
  labs(title = "Tenures and Degree Centrality", x = "Tenure Days", y = "Degree Centrality")
```



Most of the points are clustered at a low degree centrality value for both males and females, suggesting that the majority of examiners do not have a large number of connections within the network.

There are a few points with notably higher centrality, particularly among male examiners and white race. These individuals might be key nodes within the network, possibly acting as hubs of information or advice.

white examiners (blue points) are the most prevalent group. There are fewer points for Asian (red) and Hispanic (green) examiners, which may reflect their distribution within the workforce or could suggest a potential area for further investigation regarding diversity within the organization.

```
# Perform t-test
t_test_result <- t.test(degree_centrality ~ gender, data = selected_data_162)

# Print the results
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  degree_centrality by gender
## t = -16.404, df = 300, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
```

```
##  -81.48777 -64.03036
## sample estimates:
## mean in group female   mean in group male
##             26.26316             99.02222
```

The t-values here indicating a significant difference in the mean degree centrality between female and male examiners. The p-value is less than 2.2e-16.This means we reject the null hypothesis, which states there is no difference between the groups. Therefore, we have sufficient evidence to conclude that there is a statistically significant difference in the mean degree centrality between female and male examiners.