DS5110 Homework 3

Ameya Santosh Gidh

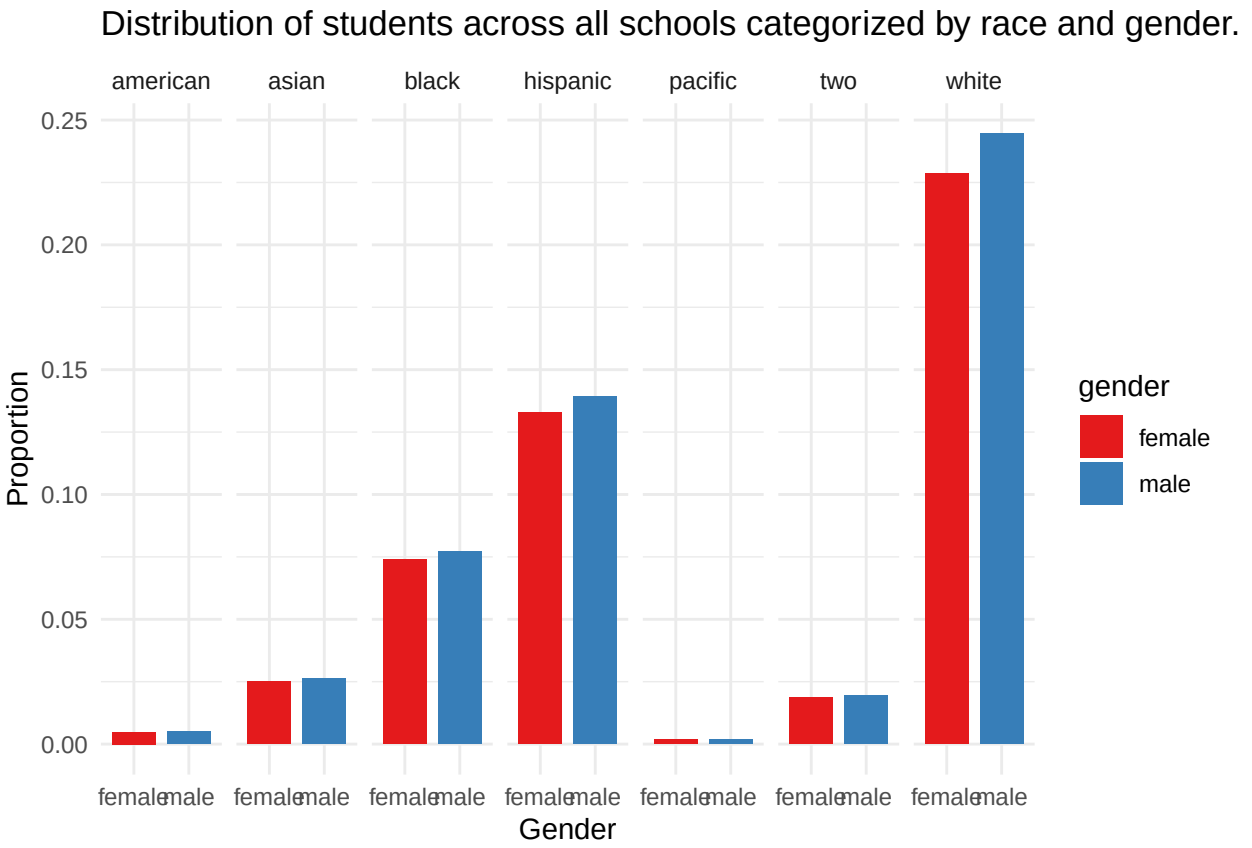
2024-02-17

Part A

Problem 1

```
# Load required libraries
suppressPackageStartupMessages(library(dplyr))
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data from csv file
env_data <- read.csv("Enrollment.csv")

# Replace negative values with NA
env_data[env_data<0] <- NA

# Summarize the data to calculate total enrollment and enrollment by race and gender
env_data <- env_data %>%
  summarize(
    total_enrollment = sum(TOT_ENR_M + TOT_ENR_F, na.rm = TRUE),
    hispanic_male_total = sum(SCH_ENR_HI_M, na.rm = TRUE),
    hispanic_female_total = sum(SCH_ENR_HI_F, na.rm = TRUE),
    american_indian_male_total = sum(SCH_ENR_AM_M, na.rm = TRUE),
    american_indian_female_total = sum(SCH_ENR_AM_F, na.rm = TRUE),
    asian_male_total = sum(SCH_ENR_AS_M, na.rm = TRUE),
    asian_female_total = sum(SCH_ENR_AS_F, na.rm = TRUE),
    pacific_islander_male_total = sum(SCH_ENR_HP_M, na.rm = TRUE),
    pacific_islander_female_total = sum(SCH_ENR_HP_F, na.rm = TRUE),
    black_male_total = sum(SCH_ENR_BL_M, na.rm = TRUE),
    black_female_total = sum(SCH_ENR_BL_F, na.rm = TRUE),
    white_male_total = sum(SCH_ENR_WH_M, na.rm = TRUE),
    white_female_total = sum(SCH_ENR_WH_F, na.rm = TRUE),
    two_or_more_male_total = sum(SCH_ENR_TR_M, na.rm = TRUE),
    two_or_more_female_total = sum(SCH_ENR_TR_F, na.rm = TRUE),
  )

# Calculate the proportion of students by race and gender
total_enrollment_prop <- env_data %>%
  mutate(
    hispanic_male = hispanic_male_total / total_enrollment,
    hispanic_female = hispanic_female_total / total_enrollment,
    american_male = american_indian_male_total / total_enrollment,
```

```
    american_female = american_indian_female_total / total_enrollment,
    asian_male = asian_male_total / total_enrollment,
    asian_female = asian_female_total / total_enrollment,
    pacific_male = pacific_islander_male_total / total_enrollment,
    pacific_female = pacific_islander_female_total / total_enrollment,
    black_male = black_male_total / total_enrollment,
    black_female = black_female_total / total_enrollment,
    white_male = white_male_total / total_enrollment,
    white_female = white_female_total / total_enrollment,
    two_male = two_or_more_male_total / total_enrollment,
    two_female = two_or_more_female_total / total_enrollment,
  ) %>%
  select(
    ends_with('male'),
    ends_with('female'))

# Transform data for plotting
total_enrollment_prop %>%
  gather(key = "race_gender", value = "proportion") %>%
  separate(race_gender, into = c("race", "gender"), sep = "_") %>%
  ggplot(aes(x = gender, y = proportion, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.8) +
  labs(x = "Gender", y = "Proportion",
       title = "Distribution of students across all schools categorized by race and gender.") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal() +
  facet_grid(. ~ race)
```



The visualization highlights White students as the most numerous across all races and genders. Conversely, Pacific Islanders and American Indians exhibit notably lower enrollment rates. Specifically, White enrollment nearly doubles that of Hispanics, the second most enrolled race. Overall, there's significant racial inequality in enrollments, while gender enrollment rates remain relatively consistent across races.

Observations: 1. White students comprise the largest proportion of enrolled students across all schools, regardless of gender. 2. Students of Native Hawaiian or Other Pacific Islander ethnicity have the smallest proportion of both male and female enrollments. 3. Gender comparison across races: - Male student populations are notably higher for White, Hispanic, and Black races. - Male enrollment slightly surpasses that of females for Asian and Two or More races. - Native American and Native Hawaiian or Other Pacific Islander races exhibit an equal distribution of male and female students.

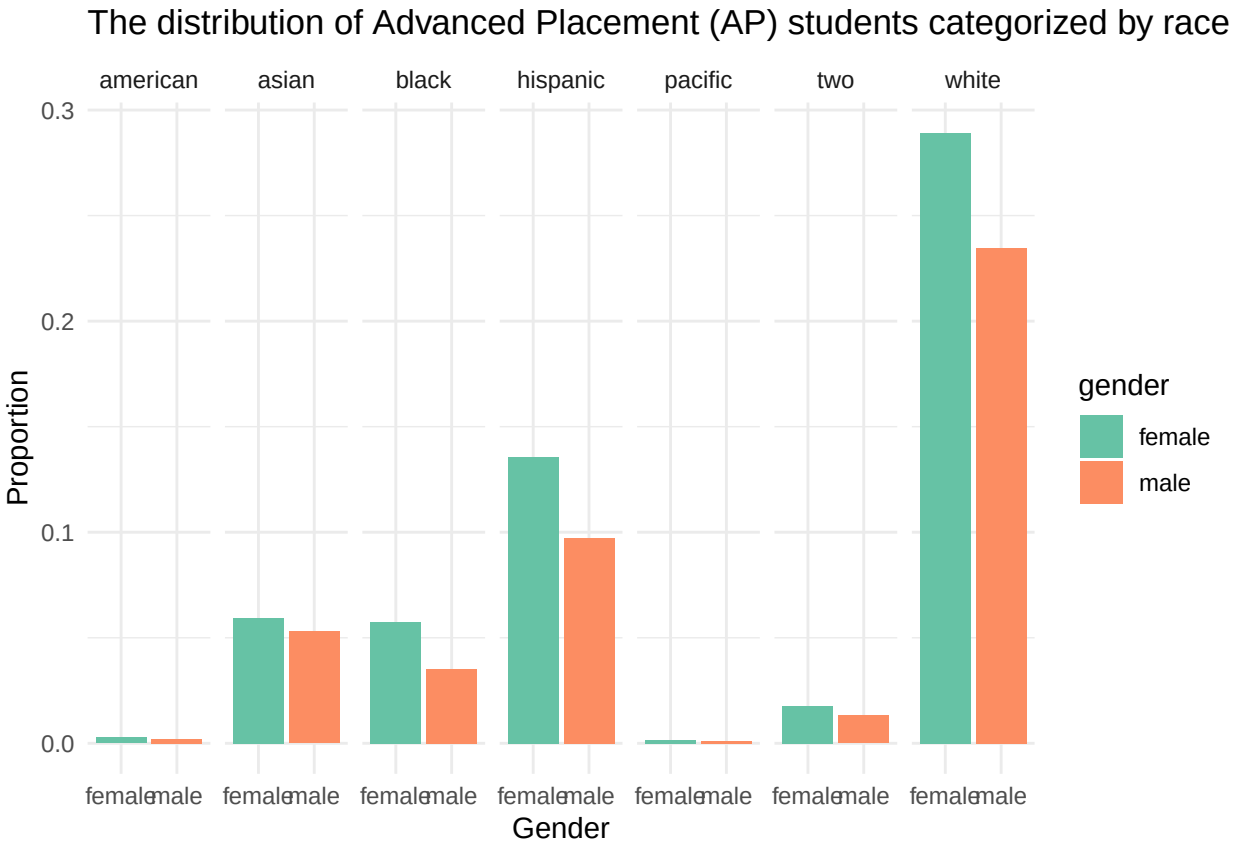
Problem 2

```r
# Load required libraries
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data from csv file
adv_placement_data <- read.csv("Advanced Placement.csv")

# Replace negative values with NA
adv_placement_data[adv_placement_data<0] <- NA

# Filter out rows where SCH_APENR_IND is not NA
adv_placement_data <- adv_placement_data[!is.na(adv_placement_data$SCH_APENR_IND),]
```

```r
# Filter out rows where SCH_APENR_IND is "Yes"
adv_placement_data <- adv_placement_data[adv_placement_data$SCH_APENR_IND == "Yes",]

# Summarize the data to calculate total enrollment and enrollment by race
# and gender for AP courses

adv_placement_data <- adv_placement_data %>%
  summarize(
    total_enrollment = sum(TOT_APENR_M + TOT_APENR_F, na.rm = TRUE),
    hispanic_male_total = sum(SCH_APENR_HI_M, na.rm = TRUE),
    hispanic_female_total = sum(SCH_APENR_HI_F, na.rm = TRUE),
    american_indian_male_total = sum(SCH_APENR_AM_M, na.rm = TRUE),
    american_indian_female_total = sum(SCH_APENR_AM_F, na.rm = TRUE),
    asian_male_total = sum(SCH_APENR_AS_M, na.rm = TRUE),
    asian_female_total = sum(SCH_APENR_AS_F, na.rm = TRUE),
    pacific_islander_male_total = sum(SCH_APENR_HP_M, na.rm = TRUE),
    pacific_islander_female_total = sum(SCH_APENR_HP_F, na.rm = TRUE),
    black_male_total = sum(SCH_APENR_BL_M, na.rm = TRUE),
    black_female_total = sum(SCH_APENR_BL_F, na.rm = TRUE),
    white_male_total = sum(SCH_APENR_WH_M, na.rm = TRUE),
    white_female_total = sum(SCH_APENR_WH_F, na.rm = TRUE),
    two_or_more_male_total = sum(SCH_APENR_TR_M, na.rm = TRUE),
    two_or_more_female_total = sum(SCH_APENR_TR_F, na.rm = TRUE),
  )

# Calculate the proportion of students by race and gender
total_enrollment_prop <- adv_placement_data %>%
  mutate(
    hispanic_male = hispanic_male_total / total_enrollment,
    hispanic_female = hispanic_female_total / total_enrollment,
    american_male = american_indian_male_total / total_enrollment,
    american_female = american_indian_female_total / total_enrollment,
    asian_male = asian_male_total / total_enrollment,
    asian_female = asian_female_total / total_enrollment,
    pacific_male = pacific_islander_male_total / total_enrollment,
    pacific_female = pacific_islander_female_total / total_enrollment,
    black_male = black_male_total / total_enrollment,
    black_female = black_female_total / total_enrollment,
    white_male = white_male_total / total_enrollment,
    white_female = white_female_total / total_enrollment,
    two_male = two_or_more_male_total / total_enrollment,
    two_female = two_or_more_female_total / total_enrollment,
  ) %>%
  select(
    ends_with('male'),
    ends_with('female'))

# Transform data for plotting
total_enrollment_prop %>%
  gather(key = "race_gender", value = "proportion") %>%
  separate(race_gender,
           into = c("race", "gender"),
           sep = "_") %>%
```

```r
  ggplot(aes(x = gender, y = proportion, fill = gender)) +
  geom_col(position = "dodge", stat = "identity") +
  labs(x = "Gender", y = "Proportion",
       title = "The distribution of Advanced Placement (AP) students categorized by race and gender acr
  scale_fill_brewer(palette = "Set2") +
  theme_minimal() +
  facet_grid(. ~ race)
```

```
## Warning in geom_col(position = "dodge", stat = "identity"): Ignoring unknown
## parameters: `stat`
```



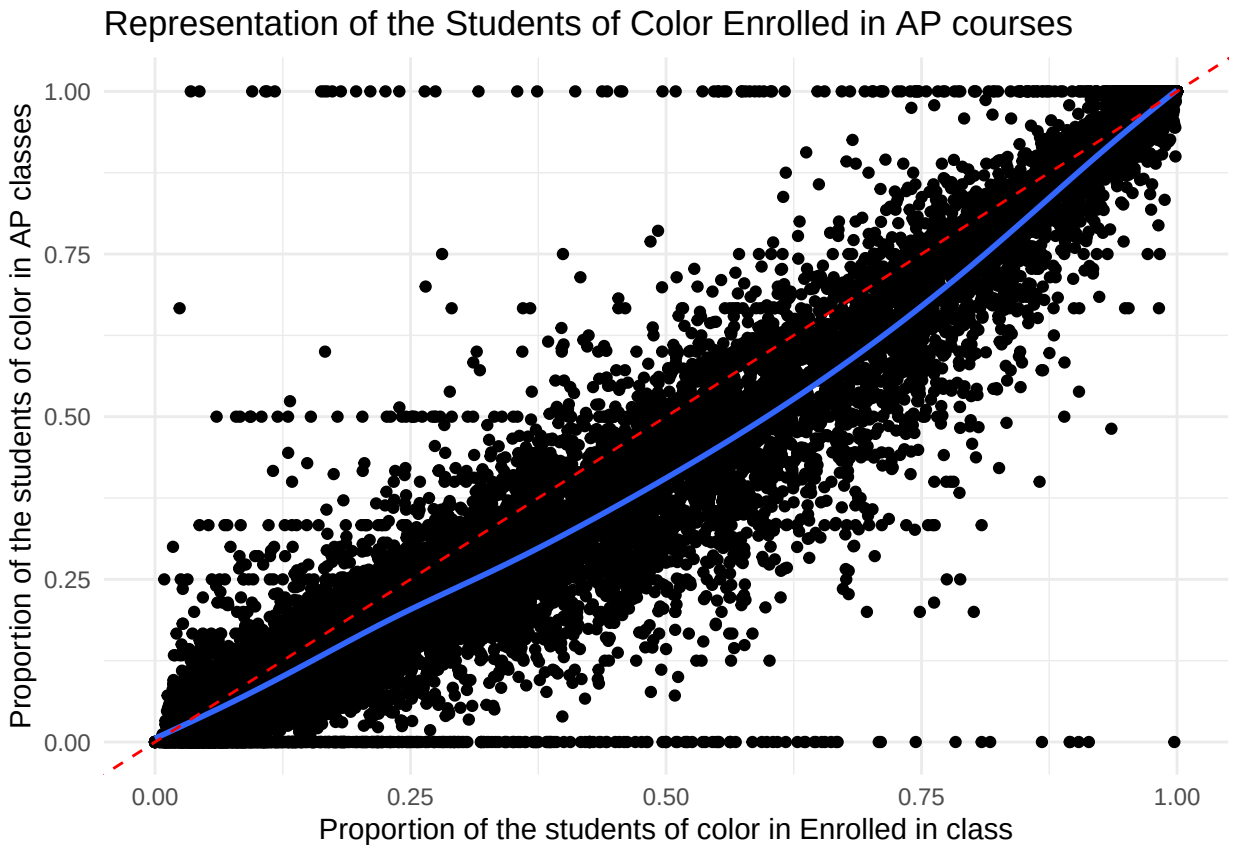
The plot is concerning because it highlights that the White race is disproportionately over represented in enrollments compared to other races. Another noteworthy finding is that for AP courses, females are either enrolled more or equally as males, regardless of race. Enrollments from American Indians and Pacific Islanders are particularly low compared to other races. There is a substantial decline in enrollments for AP courses among all other races except Whites (who, once again, have more than twice the enrollment rate of the second-highest enrolled race in AP courses).

Observations: 1. Female students outnumber male students in the AP program across all racial groups. 2. White students comprise the majority of both male and female students. 3. The smallest number of male and female students are from the 'Native Hawaiian or Other Pacific Islander' race. 4. Comparing to the previous observation, there was a greater male dominance across all races in the first quarter, whereas in the second quarter, there's a higher percentage of females across all races. 5. There are variations in distribution across races: Asian and Black populations are more represented in AP programs.

Problem 3

```
# Load the packages
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data from csv file
adv_placement_data <- read.csv("Advanced Placement.csv")

# Replace negative values with NA
adv_placement_data[adv_placement_data<0] <- NA

# Filter out rows where SCH_APENR_IND is not NA
adv_placement_data <- adv_placement_data[!is.na(adv_placement_data$SCH_APENR_IND),]

# Filter out rows where SCH_APENR_IND is "Yes"
adv_placement_data <- adv_placement_data[adv_placement_data$SCH_APENR_IND == "Yes",]

# Select columns of students of color from the dataset
adv_placement_data <- adv_placement_data %>%
  select(COMBOKEY, TOT_APENR_M, TOT_APENR_F, SCH_APENR_HI_M, SCH_APENR_HI_F,
         SCH_APENR_AM_M, SCH_APENR_AM_F, SCH_APENR_AS_M, SCH_APENR_AS_F,
         SCH_APENR_HP_M, SCH_APENR_HP_F, SCH_APENR_BL_M, SCH_APENR_BL_F,
         SCH_APENR_TR_M, SCH_APENR_TR_F)

# Group the data by "COMBOKEY" and calculate the total number of students,
# total number of students of color, and proportion of students of color in AP classes
adv_placement_data <- adv_placement_data %>%
  group_by(COMBOKEY) %>%
  summarise(total_students_ap = sum(TOT_APENR_M, TOT_APENR_F, na.rm = TRUE),
            total_students_of_color_ap = sum(SCH_APENR_HI_M, SCH_APENR_HI_F,
                                             SCH_APENR_AM_M, SCH_APENR_AM_F,
                                             SCH_APENR_AS_M, SCH_APENR_AS_F,
                                             SCH_APENR_HP_M, SCH_APENR_HP_F,
                                             SCH_APENR_BL_M, SCH_APENR_BL_F,
                                             SCH_APENR_TR_M, SCH_APENR_TR_F,
                                             na.rm = TRUE),
            prop_students_of_color_ap = total_students_of_color_ap/total_students_ap)

# The code removes any rows with missing values from the dataset
adv_placement_data <- na.omit(adv_placement_data)

# Load data from csv file
env_data <- read.csv("Enrollment.csv")

# Replace negative values with NA
env_data[env_data < 0] <- NA

# Filter out the schools where AP courses are conducted
env_data <- env_data[env_data$COMBOKEY %in% adv_placement_data$COMBOKEY, ]

# Select columns of students of color from the dataset
env_data <- env_data %>%
```

```r
  select(COMBOKEY, TOT_ENR_M, TOT_ENR_F, SCH_ENR_HI_M, SCH_ENR_HI_F, SCH_ENR_AM_M,
         SCH_ENR_AM_F, SCH_ENR_AS_M, SCH_ENR_AS_F, SCH_ENR_HP_M, SCH_ENR_HP_F,
         SCH_ENR_BL_M, SCH_ENR_BL_F, SCH_ENR_TR_M, SCH_ENR_TR_F)

# Group the data by "COMBOKEY" and calculates the total number of students,
# total number of students of color, and proportion of students of color in enrolled classes

env_data <- env_data %>%
  group_by(COMBOKEY) %>%
  summarise(total_students = sum(TOT_ENR_M, TOT_ENR_F, na.rm = TRUE),
            total_students_of_color = sum(SCH_ENR_HI_M, SCH_ENR_HI_F,
                                          SCH_ENR_AM_M, SCH_ENR_AM_F,
                                          SCH_ENR_AS_M, SCH_ENR_AS_F,
                                          SCH_ENR_HP_M, SCH_ENR_HP_F,
                                          SCH_ENR_BL_M, SCH_ENR_BL_F,
                                          SCH_ENR_TR_M, SCH_ENR_TR_F, na.rm = TRUE),
            prop_students_of_color = total_students_of_color/total_students)

# Filter out the COMBOKEY and Proportion of students of color enrolled overall
enrollment_prop_students_of_color <- env_data %>%
  select(COMBOKEY, prop_students_of_color)

# Filter out the COMBOKEY and Proportion of students of color enrolled in AP courses
ap_data_prop_students_of_color_ap <- adv_placement_data %>%
  select(COMBOKEY, prop_students_of_color_ap)

# Merge the data on COMBOKEY
merged_data <- merge(enrollment_prop_students_of_color,
                     ap_data_prop_students_of_color_ap, by = "COMBOKEY")

# Plot the distribution graph
ggplot(merged_data, aes(x = prop_students_of_color, y = prop_students_of_color_ap)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'gam', formula = y ~ s(x, bs = "cs")) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Proportion of the students of color in Enrolled in class",
       y = "Proportion of the students of color in AP classes",
       title = "Representation of the Students of Color Enrolled in AP courses") +
  theme_minimal()
```



The plot illustrates the relationship between the proportion of students of color in Advanced Placement (AP) courses and the overall enrolled population. While there's generally a positive correlation between the two, many schools show a lower proportion of students of color in AP courses, suggesting under-representation. The presence of a smoothing line emphasizes this trend, underscoring the need to address the under-representation and ensure equitable access to advanced coursework. Variations in the representation of students of color across schools are evident from points above or below the red line. Some points near the extremes of the x-axis may indicate significantly higher or lower proportions of students of color in enrolled classes compared to AP courses. It's crucial to address these disparities and promote equal access and opportunities for all students.

Observations: 1. Regarding the question of whether students of color are often underrepresented in AP classes: - The large number of points below the reference line (the intercept) indicates that students of color are typically underrepresented in AP classes. - There exists a positive correlation between the proportion of students of color across all schools and the proportion of non-white students enrolled in at least one AP class.

Part B

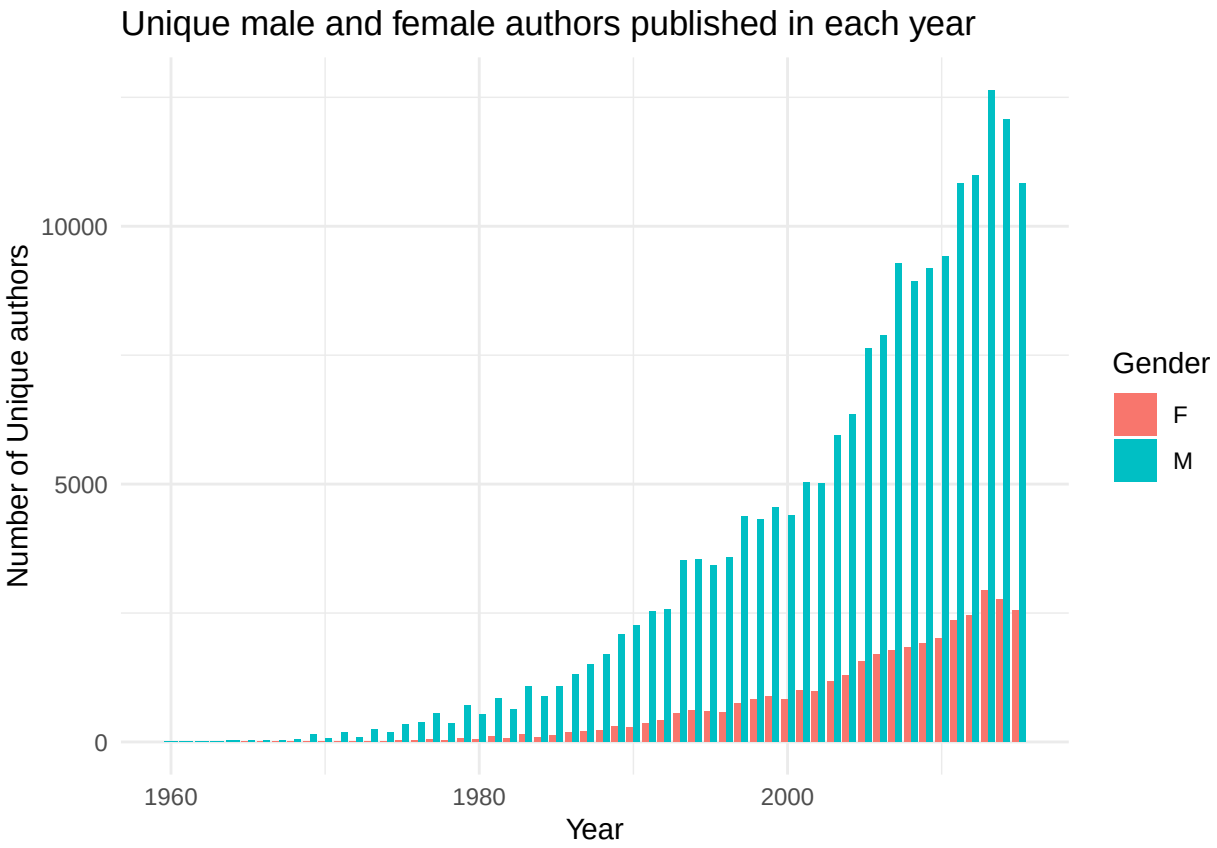
Problem 4

```
# Load required packages
library(RSQLite)
library(tidyr)
library(dplyr)
library(ggplot2)

# Connect to the SQLite database file
db <- dbConnect(RSQLite::SQLite(), "dblp.db")
```

```r
tables <- dbListTables(db)
tables
```

```
## [1] "affiliation"       "affiliation_coord" "authors"
## [4] "editors"           "genauth_old"       "genedit"
## [7] "general"
```

```r
# query_result <- dbGetQuery(db, "SELECT * FROM general")
# head(query_result)

# query_result <- dbGetQuery(db, "SELECT * FROM authors")
# head(query_result)

# Filter out non-male and non-female authors with prediction probability less than 0.9
query <- "SELECT * FROM general JOIN authors ON general.k = authors.k"
df <- dbGetQuery(db, query)
df <- df[df$gender %in% c("M", "F") & df$prob >= 0.9,]

# Remove the "k" column as its duplicate and not needed
df <- select(df, -k)

# Disconnect from the database
dbDisconnect(db)

# Aggregate the number of distinct authors by year and gender
author_counts <- aggregate(name ~ year + gender, data = df, FUN = function(x)
  length(unique(x)))

# Create a bar plot of the number of distinct male and female authors published each year
ggplot(author_counts, aes(x = year, y = name, fill = gender)) +
  geom_col(position = "dodge") +
  xlab("Year") +
  ylab("Number of Unique authors") +
  ggtitle("Unique male and female authors published in each year") +
  labs(fill = "Gender") +
  theme_minimal()
```



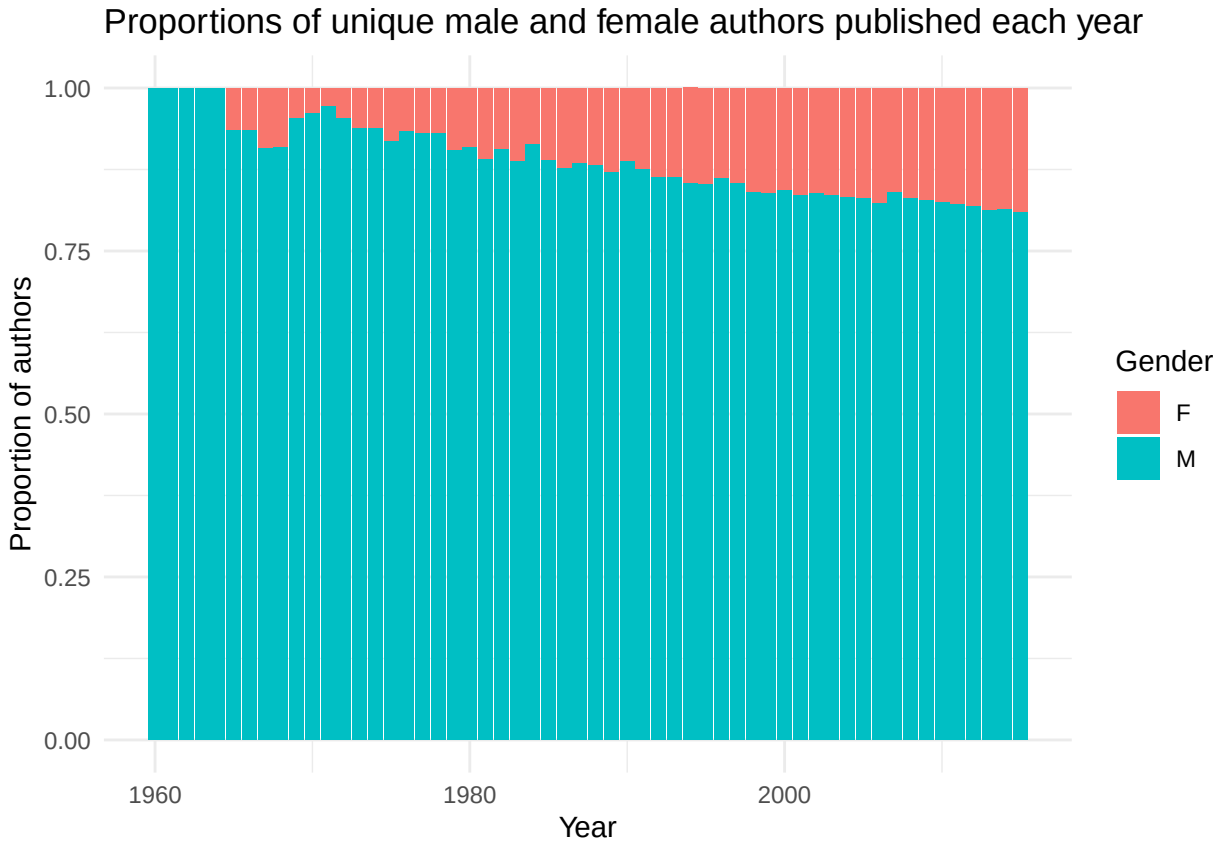
The visualization displays how the publication of authors by gender has evolved over time, indicating a steady increase in the total number of authors. However, it also emphasizes a widening gap between male and female authors, with males consistently surpassing females in numbers. This suggests possible gender biases or barriers that impede women from publishing at the same pace as men, emphasizing the importance of tackling these challenges to ensure equal opportunities for researchers of all genders.

Observations: 1. The count of unique male authors consistently surpasses that of unique female authors throughout the years. 2. There's a noticeable upward trend in the total number of authors over time. 3. The count of female authors barely exceeds 2500.

Problem 5

```r
# Load required packages
library(RSQLite)
library(tidyr)
library(dplyr)
library(ggplot2)

# Connect to the database
db <- dbConnect(RSQLite::SQLite(), "dblp.db")

# Create a query to join the "general" and "authors" tables
query <- "SELECT * FROM general JOIN authors ON general.k = authors.k"

# Retrieve the query result into a dataframe
df <- dbGetQuery(db, query)

# Filter out non-male and non-female authors with prediction probability less than 0.9
```

```r
df <- df[df$gender %in% c("M", "F") & df$prob >= 0.9,]

# Remove the "k" column
df <- select(df, -k)

# Disconnect from the database
dbDisconnect(db)

# Aggregate the number of distinct authors by year and gender
author_counts <- aggregate(name ~ year + gender, data = df, FUN = function(x)
  length(unique(x)))

# Group the author_counts by year and gender, calculate the proportion of
# authors and select only relevant columns
author_props <- author_counts %>%
  group_by(year) %>%
  mutate(prop = name/sum(name)) %>%
  select(year, gender, prop)

# Plot the proportion of distinct male and female authors published each year
ggplot(author_props, aes(x = year, y = prop, fill = gender)) +
  geom_col(position = "stack") +
  xlab("Year") +
  ylab("Proportion of authors") +
  ggtitle("Proportions of unique male and female authors published each year") +
  labs(fill = "Gender") +
  theme_minimal()
```



The visualization displays the proportions of unique male and female authors published annually. It reveals a gradual rise in the proportion of female authors over time, although males still hold a higher proportion. However, the gap between male and female authorship appears to be narrowing gradually, suggesting that initiatives to foster gender equality in authorship may be yielding positive results. The visualization underscores the issue of gender disparity in authorship and underscores the ongoing need for initiatives promoting gender diversity in academic publishing.

Observations: 1. The ratio of male authors has consistently exceeded that of female authors over the years. 2. When the proportions for each year are totaled, the sum equals 1. 3. In the initial years, there are no female authors at all, resulting in a male author proportion of 1 for those years.